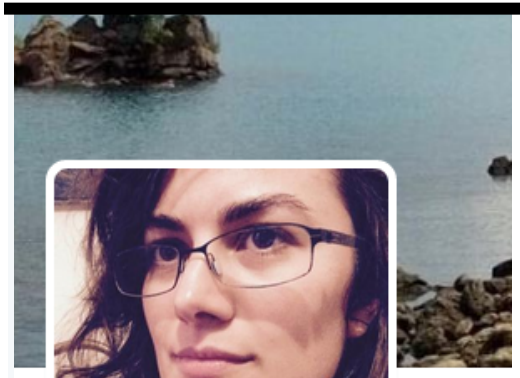
Visualizing Public Health Data

Anamaria Crisan, MSc
PhD student with Drs. Jennifer Gardy & Tamara Munzner
UBC School of Population and Public Health

Why am I giving this talk?



Ana Crisan
@amcrisan

PhD student with @jennifergardy and @tamaramunzner. I study the confluence of heterogenous clinical data streams using #stats, #bioinformatics, and #infovis

cs.ubc.ca/~acrisan

Master of Science
(Bioinformatics)

PhD
(“Computational Public Health”)

2008 — 2010 — 2013 — 2015

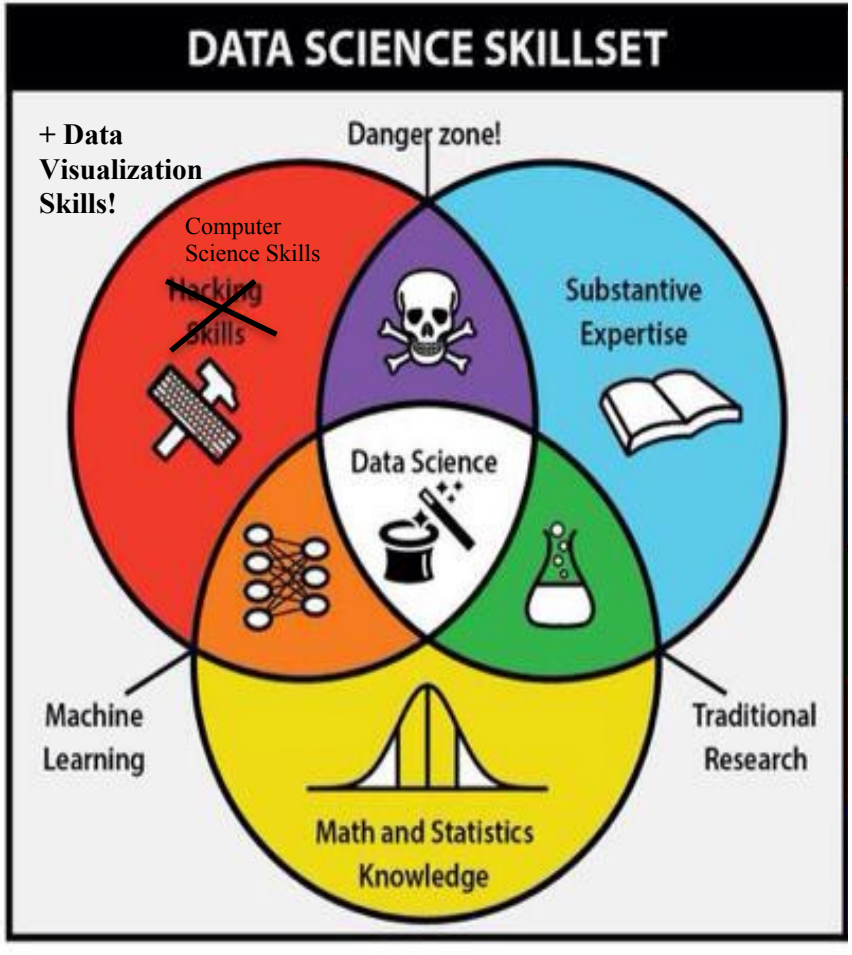
British Columbia Centre for Disease Control

GenomeDx Biosciences

@amcrisan

http://cs.ubc.ca/~acrisan

I'm not an artist. I'm a data analyst.



http://blog.framed.io/

Disclaimer

I'll be talking about a project I worked on while employed at GenomeDx Biosciences.

Everything I am presenting is publically available, but this doesn't mean that I endorse their products or the products of their competitors.

Furthermore, I am relaying high level details of my own thought process during and after this project, not the thoughts of others at the organization.

Eventually I had Explain my Work to Experts with Different Backgrounds

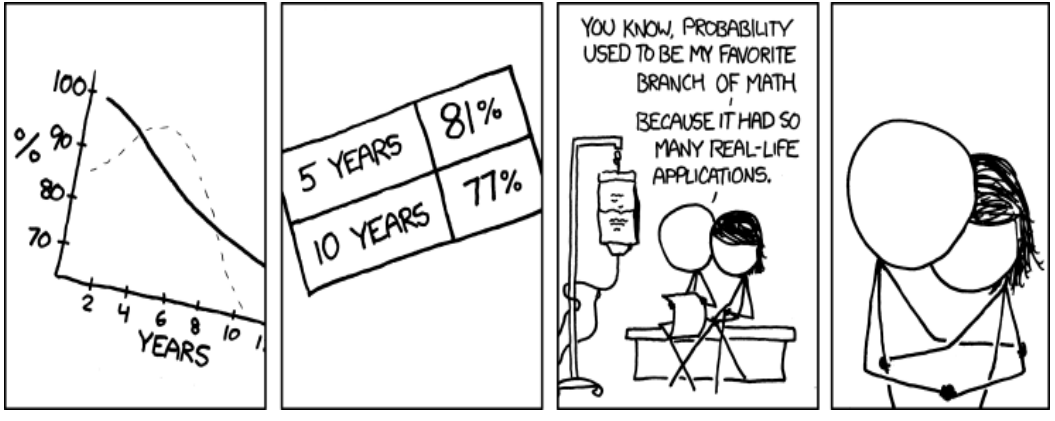
I often used data visualization to explain the results of data mining and statistical techniques

But one day I got tasked with a rather challenging problem…

The Question:

The task: We had developed a genomic biomarker panel to assess a man's risk of metastatic prostate cancer following prostatectomy

How do we communicate "risk"?



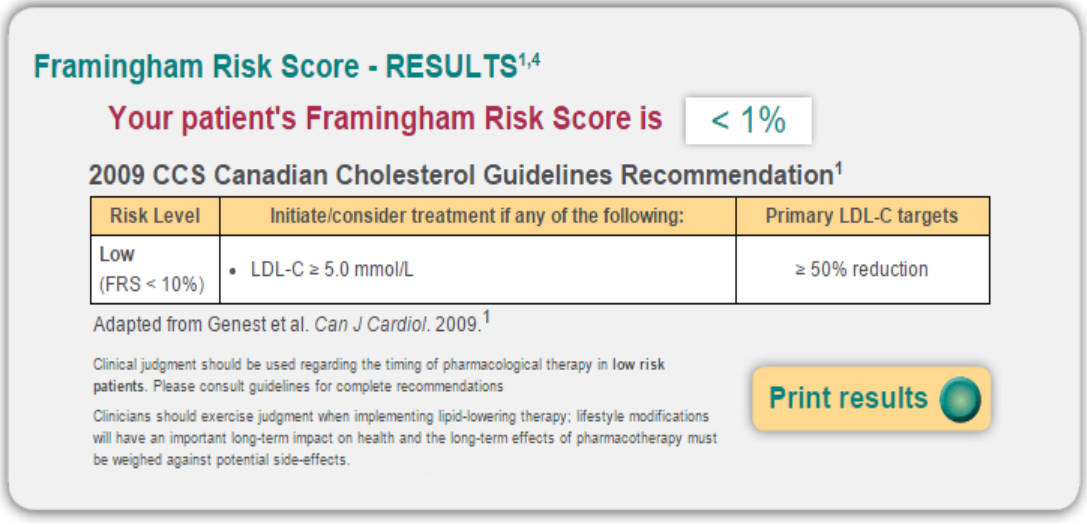
XKCD Comic #881

I wanted to take more ownership of the question “how do we communicate risk?”

I wanted to take more ownership of the question “how do we communicate risk?”

There wasn’t a simple answer

Just show a Number …



Framingham Risk Score - RESULTS[1,4]

Your patient's Framingham Risk Score is < 1%

2009 CCS Canadian Cholesterol Guidelines Recommendation[1]

Risk Level	Initiate/consider treatment if any of the following:	Primary LDL-C targets
Low (FRS < 10%)	• LDL-C ≥ 5.0 mmol/L	≥ 50% reduction

Adapted from Genest et al. *Can J Cardiol*. 2009.[1]

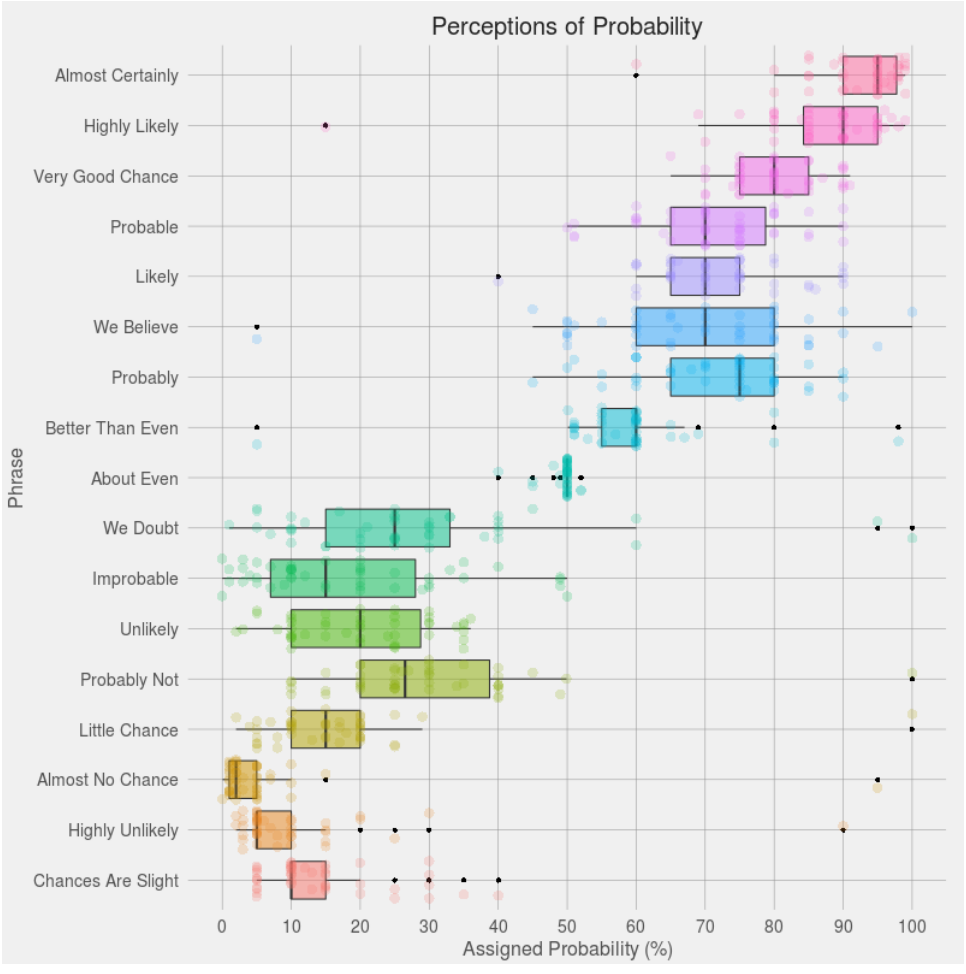
Clinical judgment should be used regarding the timing of pharmacological therapy in **low risk patients**. Please consult guidelines for complete recommendations

Clinicians should exercise judgment when implementing lipid-lowering therapy; lifestyle modifications will have an important long-term impact on health and the long-term effects of pharmacotherapy must be weighed against potential side-effects.

Print results

http://bit.ly/1Knrj19

Is a Data Visualization really Necessary?



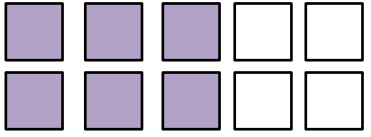
http://bit.ly/1FxtT2z

Evidence from Risk Communication Literature

(difficult to understand) *(easier to understand)*

Probability < Frequency < Visualization

60% 6 in 10



Numeracy : the ability to reason with numbers

Individuals with **low numeracy** have a difficulty interpreting numbers and probabilities

Visualizations can help people with low numeracy make sense of data,

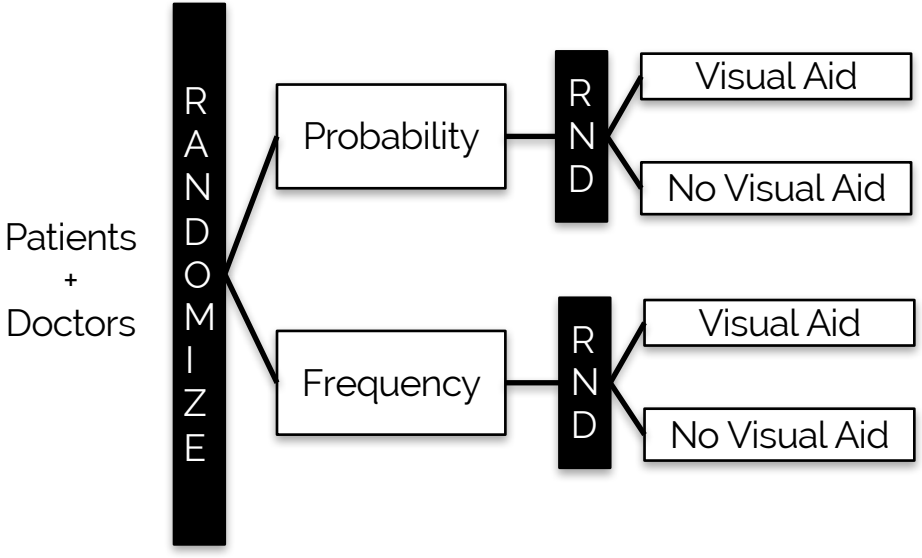
But, there is some evidence that **low numeracy** affects reasoning with graphs as well.

Whiting *et. al* (2015) "How well do health professionals interpret diagnostic information? A systematic review"

Example : Data Visualization in Shared decision Making

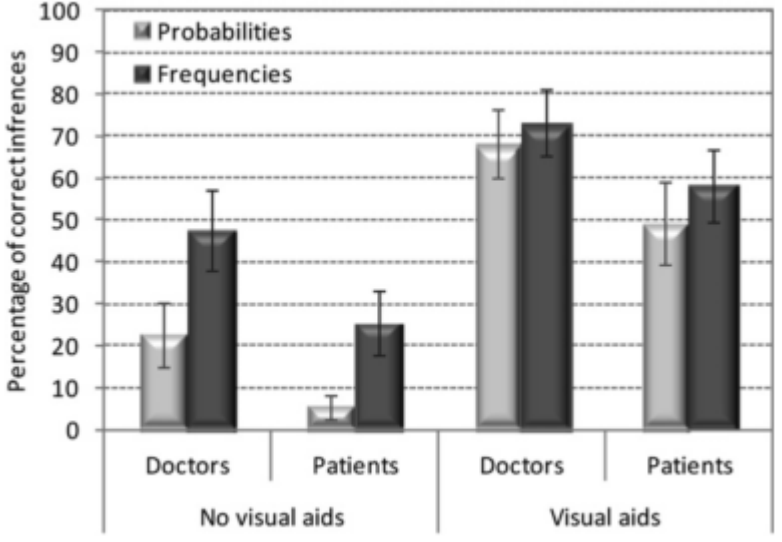
STUDY DESIGN

Quasi-randomized trial with four conditions
Outcome : correctly calculating the risk (essentially a math test)



RESULTS

Visualization improved comprehension of both doctors and patients
Visualization improved concordance between doctors and patients

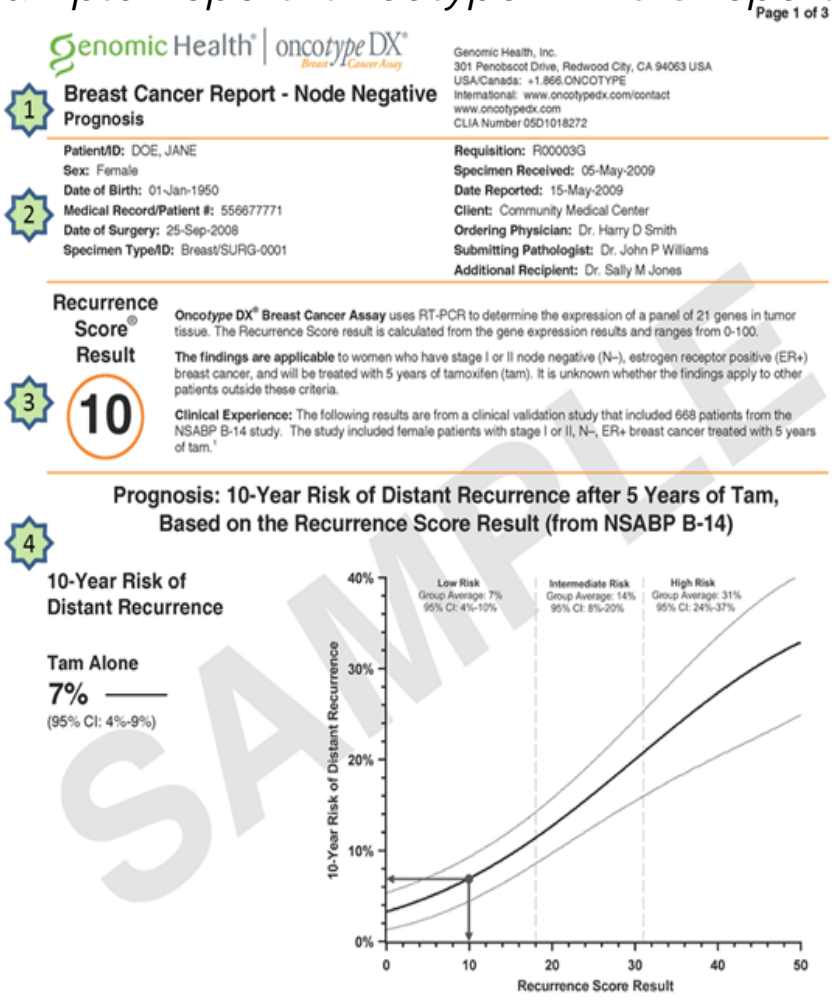


Garcia-Retamero *et. al* (2013) "Visual representation of statistical information improves diagnostic inferences in doctors and their patients"

Yes! Data visualization was more than a “nice to have”!

Show a Number and a Picture

Example Report: OncotypeDx DCIS report



Page 1 of 3

Genomic Health® | oncotype DX® Breast Cancer Assay

Genomic Health, Inc.
301 Penobscot Drive, Redwood City, CA 94063 USA
USA/Canada: +1.866.ONCOTYPE
International: www.oncotypedx.com/contact
www.oncotypedx.com
CLIA Number 05D1018272

1 **Breast Cancer Report - Node Negative Prognosis**

2

Patient/ID: DOE, JANE
Sex: Female
Date of Birth: 01-Jan-1950
Medical Record/Patient #: 556677771
Date of Surgery: 25-Sep-2008
Specimen Type/ID: Breast/SURG-0001

Requisition: R00003G
Specimen Received: 05-May-2009
Date Reported: 15-May-2009
Client: Community Medical Center
Ordering Physician: Dr. Harry D Smith
Submitting Pathologist: Dr. John P Williams
Additional Recipient: Dr. Sally M Jones

3 **Recurrence Score® Result**

10

***Oncotype DX*® Breast Cancer Assay** uses RT-PCR to determine the expression of a panel of 21 genes in tumor tissue. The Recurrence Score result is calculated from the gene expression results and ranges from 0-100.

The findings are applicable to women who have stage I or II node negative (N–), estrogen receptor positive (ER+) breast cancer, and will be treated with 5 years of tamoxifen (tam). It is unknown whether the findings apply to other patients outside these criteria.

Clinical Experience: The following results are from a clinical validation study that included 668 patients from the NSABP B-14 study. The study included female patients with stage I or II, N–, ER+ breast cancer treated with 5 years of tam.[1]

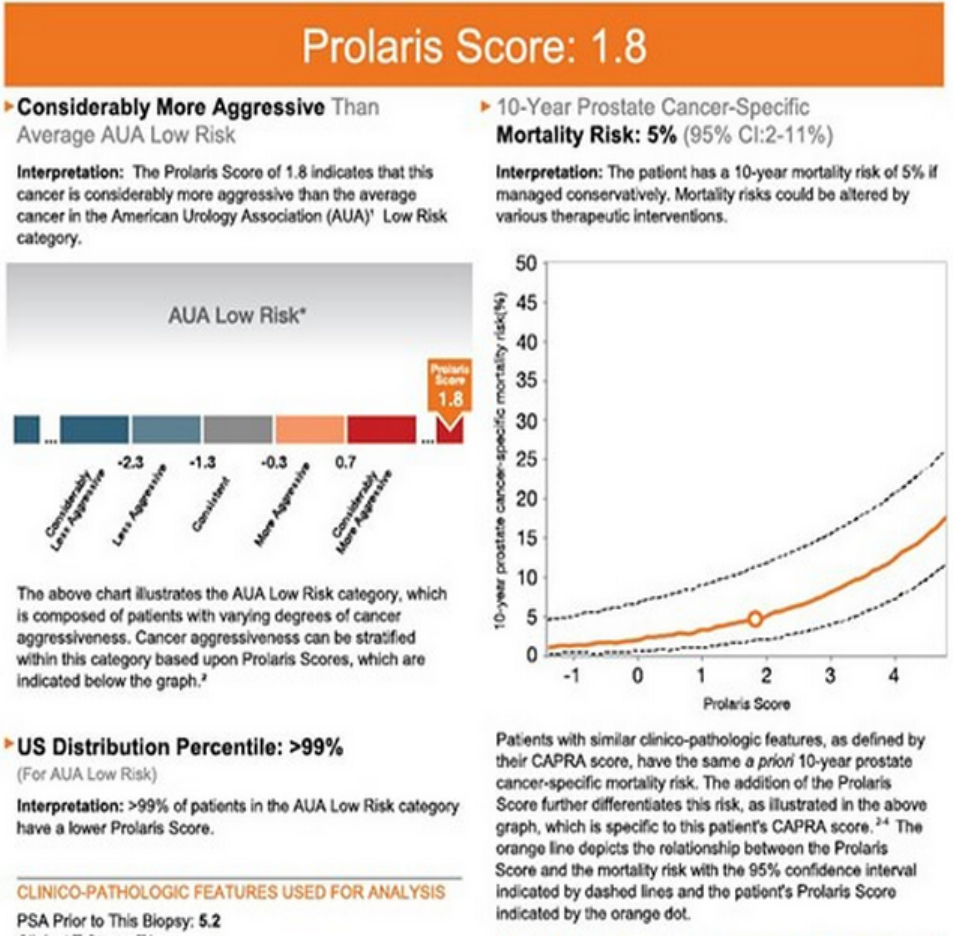
4 **Prognosis: 10-Year Risk of Distant Recurrence after 5 Years of Tam, Based on the Recurrence Score Result (from NSABP B-14)**

10-Year Risk of Distant Recurrence

Tam Alone
7%
(95% CI: 4%-9%)

Show a Number and a Picture

Example Report: Myriad Prolaris Prostate Cancer Test Report



Prolaris Score: 1.8

Considerably More Aggressive Than Average AUA Low Risk

Interpretation: The Prolaris Score of 1.8 indicates that this cancer is considerably more aggressive than the average cancer in the American Urology Association (AUA)[1] Low Risk category.

AUA Low Risk*

Prolaris Score 1.8

-2.3 -1.3 -0.3 0.7

Considerably Less Aggressive | Less Aggressive | Consistent | More Aggressive | Considerably More Aggressive

The above chart illustrates the AUA Low Risk category, which is composed of patients with varying degrees of cancer aggressiveness. Cancer aggressiveness can be stratified within this category based upon Prolaris Scores, which are indicated below the graph.[2]

US Distribution Percentile: >99%

(For AUA Low Risk)

Interpretation: >99% of patients in the AUA Low Risk category have a lower Prolaris Score.

CLINICO-PATHOLOGIC FEATURES USED FOR ANALYSIS

PSA Prior to This Biopsy: **5.2**

10-Year Prostate Cancer-Specific **Mortality Risk: 5%** (95% CI:2-11%)

Interpretation: The patient has a 10-year mortality risk of 5% if managed conservatively. Mortality risks could be altered by various therapeutic interventions.

10-year prostate cancer-specific mortality risk(%)

50 45 40 35 30 25 20 15 10 5 0

-1 0 1 2 3 4

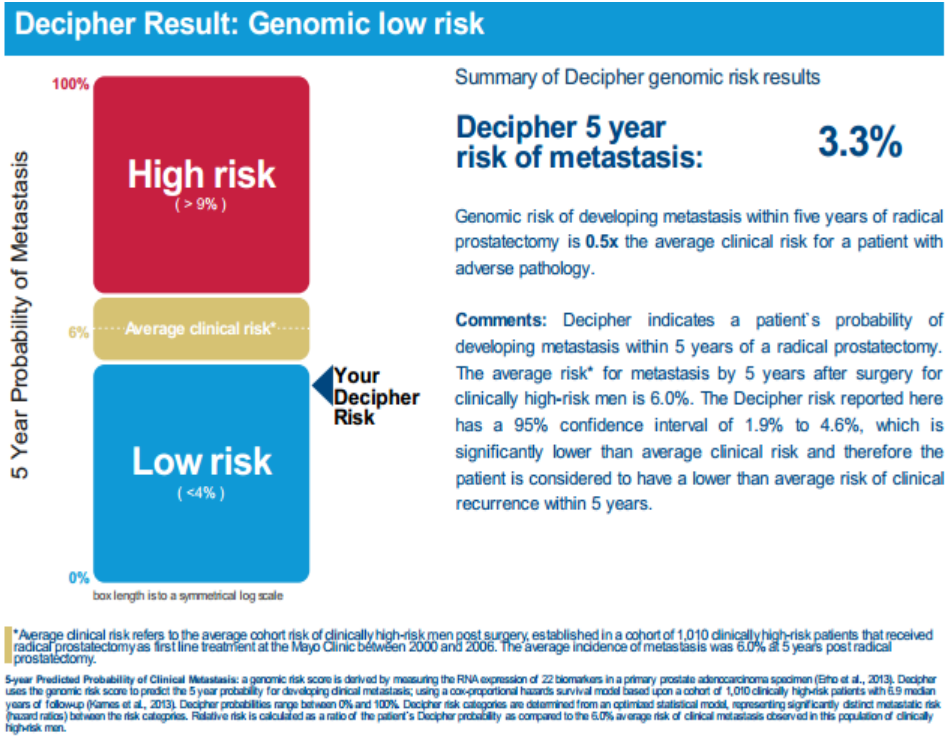
Prolaris Score

Patients with similar clinico-pathologic features, as defined by their CAPRA score, have the same *a priori* 10-year prostate cancer-specific mortality risk. The addition of the Prolaris Score further differentiates this risk, as illustrated in the above graph, which is specific to this patient's CAPRA score.[3,4] The orange line depicts the relationship between the Prolaris Score and the mortality risk with the 95% confidence interval indicated by dashed lines and the patient's Prolaris Score indicated by the orange dot.

Show a Number and a Picture

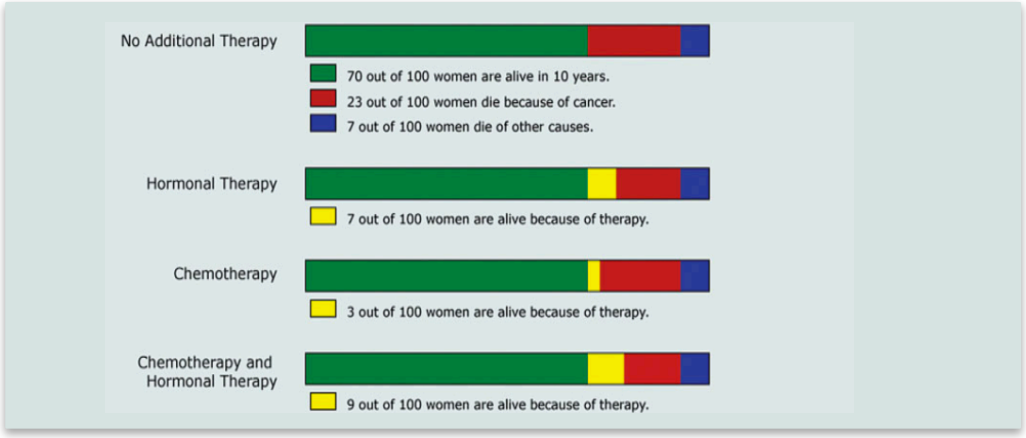
Primary population: Men, who are susceptible to red-green colour blindness

Example Report: Decipher Prostate Cancer Test Report



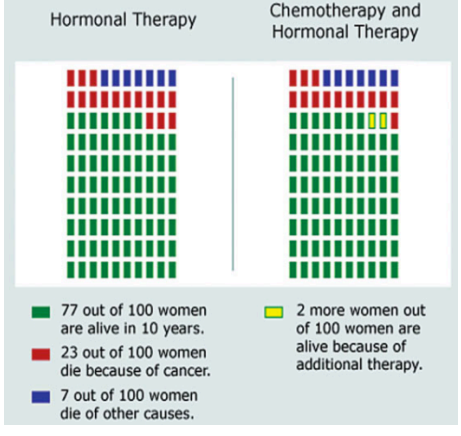
Example : Deciding upon an Intervention

Baseline Visualization

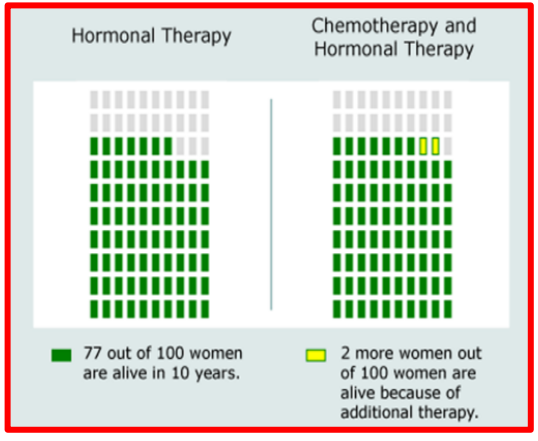


Helping breast cancer patients decide between multiple treatment options.

Alternative 1



Alternative 2

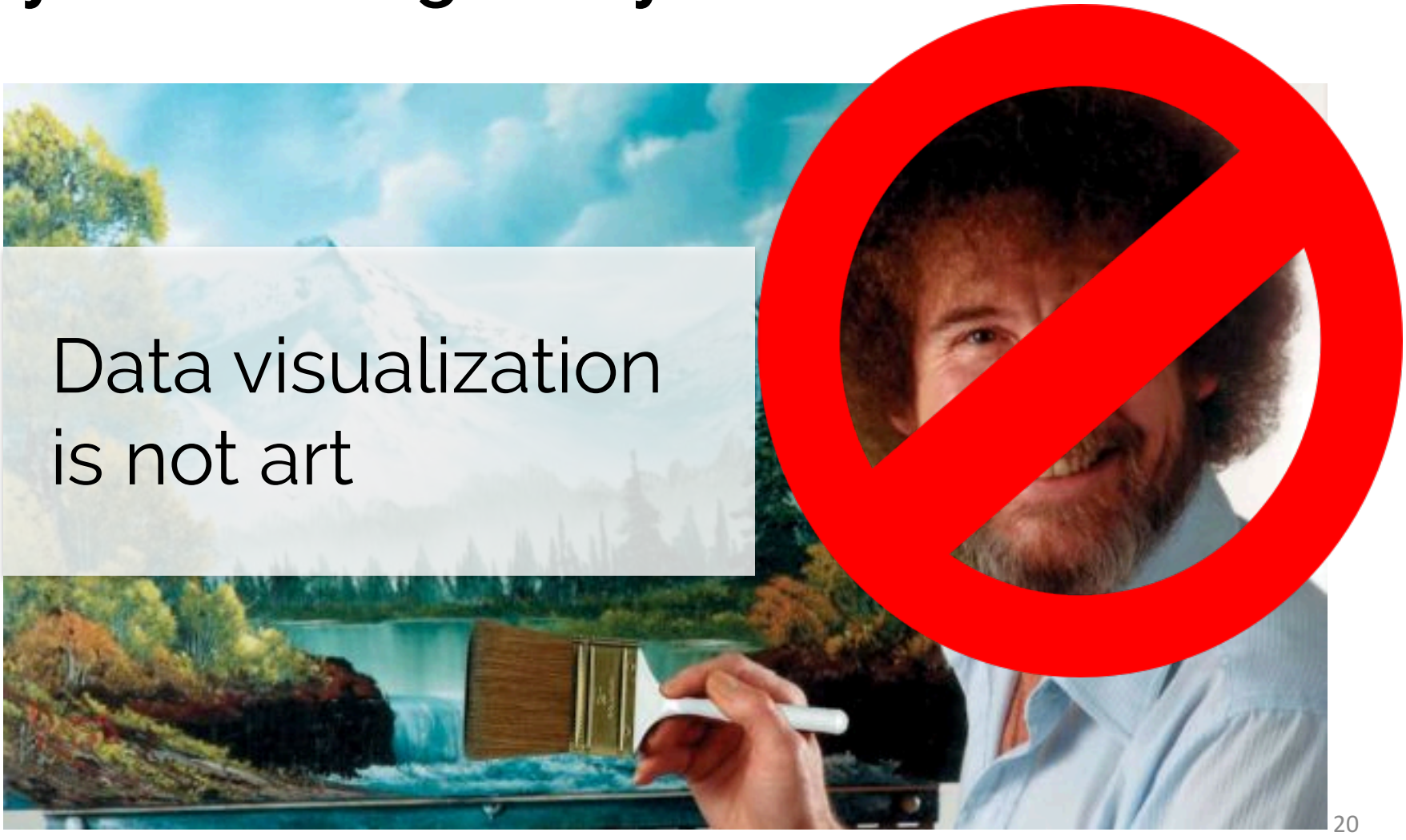


Zikmund-Fisher (2013). A demonstration of "less can be more" in risk graphics.

Zikmund-Fisher (2008). Improving understanding of adjuvant therapy options by using simpler risk graphics

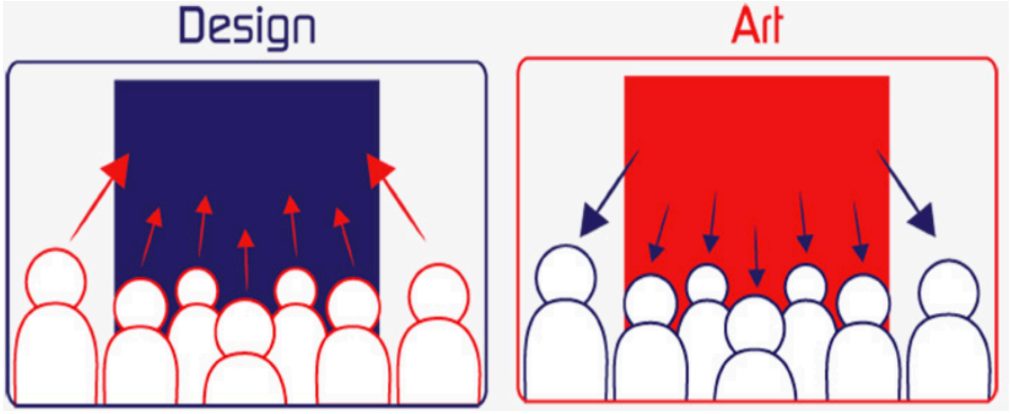
SO... what is data visualization?

Beyond Building Pretty & Cool Visualizations



Beyond Building Pretty & Cool Visualizations

Defining Data Visualization



Design **Art**

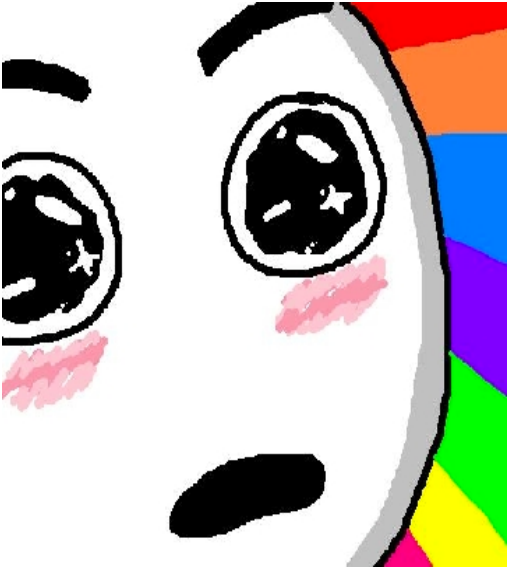
Data Visualization

(I argue data visualization is much more about design)

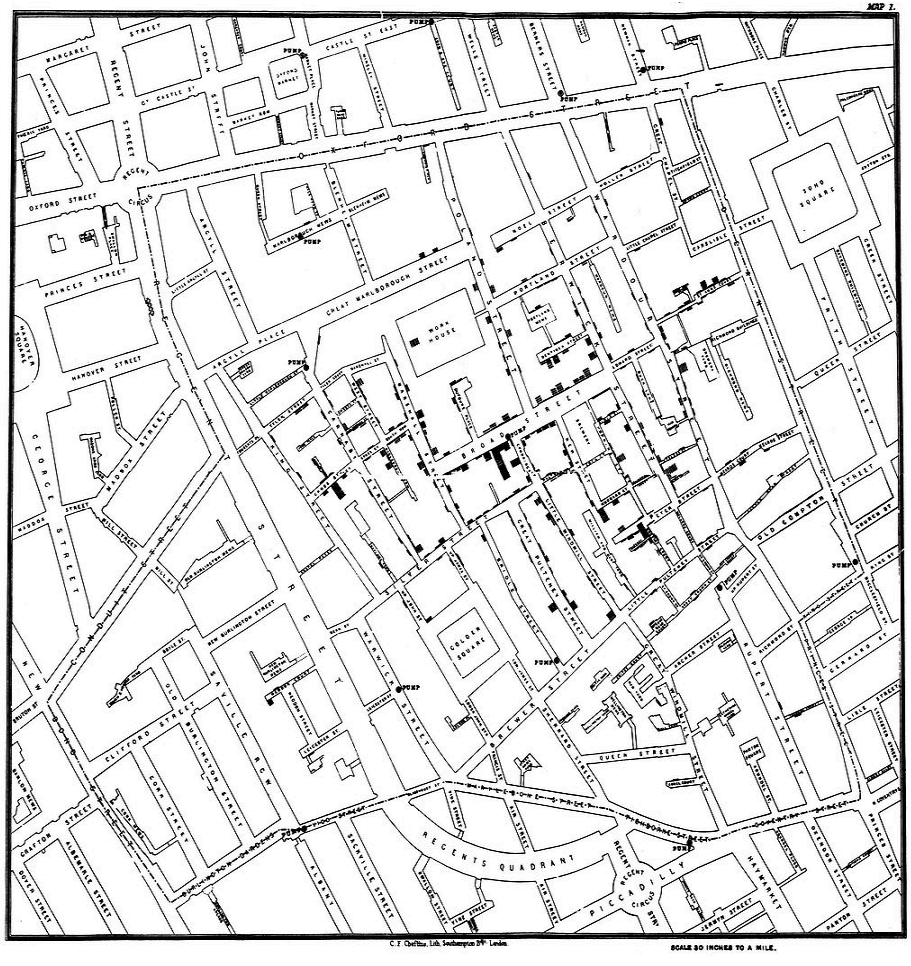
Ideas taken from @rachelbinx's 2016 OpenVis talk
And http://featureguru.com/art-vs-design.html

BUT WAIT!

There's more to data visualization than simply communicating numerical data



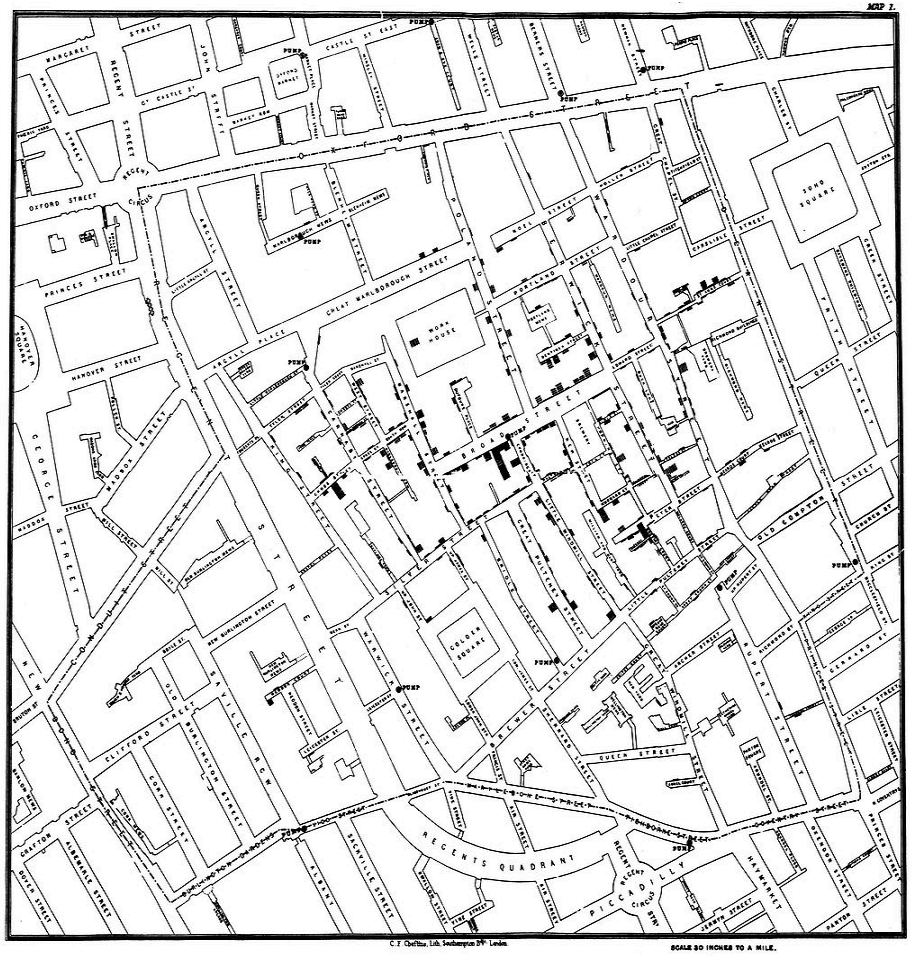
Example : Hypothesis Generation



Allowed John Snow to form the hypothesis of what may be leading to the cholera outbreak

John Snow's Visualization of the 1854 Cholera Outbreak

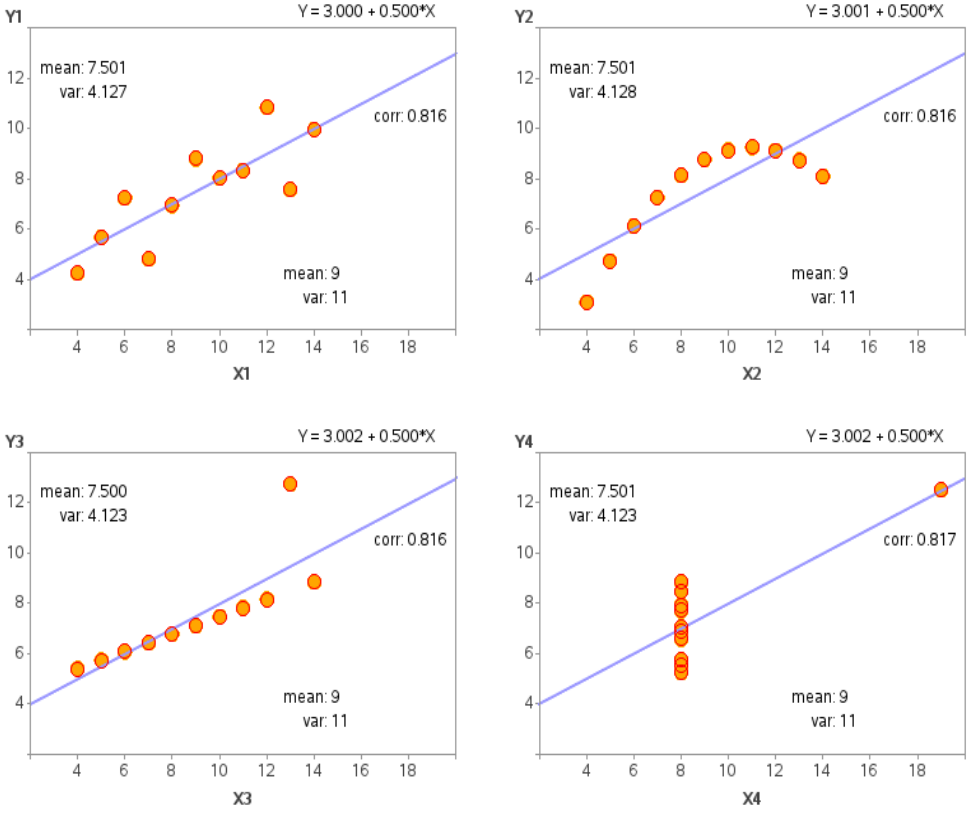
Example : Hypothesis Generation



Allowed John Snow to form the hypothesis of what may be leading to the cholera outbreak

John Snow's Visualization of the 1854 Cholera Outbreak

Example : Checking Assumptions of Statistical Models



Anscombe's quartet, four datasets that have near identical descriptive statistics but that look very different when visualized.

Anscombe, F. (1973) "Graphs in Statistical Analysis"

Data visualization has long complemented applied statistical practices. Consider Tukey's classic "Exploratory Data Analysis", which is rife with suggestions for how to visualization data.

So what should be think about when designing data visualizations?

A Data visualization in 3 Questions:

Why? (Motivation)

Why do you need to visualize data?

What? (Data)

What kind of data is being visualized?

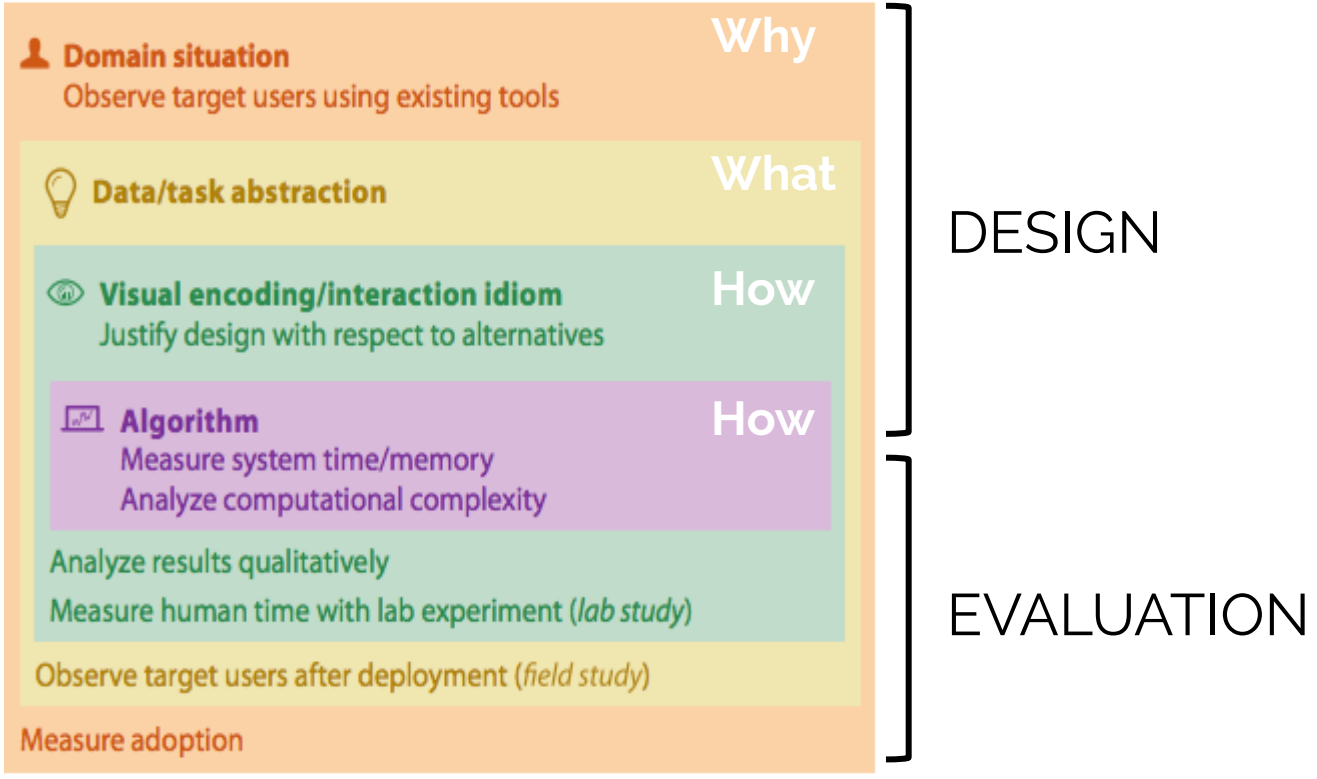
How? (Visual and Interaction Design)

How is data being visualized?

A Data visualization in 3 Questions:

Design	Evaluation
Why?	Does the visualization solve a relevant problem?
What?	Are you using the right data, or deriving the right data?
How?	Are the visual and interactive design choices appropriate?

Steps to Design and Evaluate a Data Visualization



Munzner (2014) "Visualization Analysis and Design"

Steps to Design and Evaluate a Data Visualization

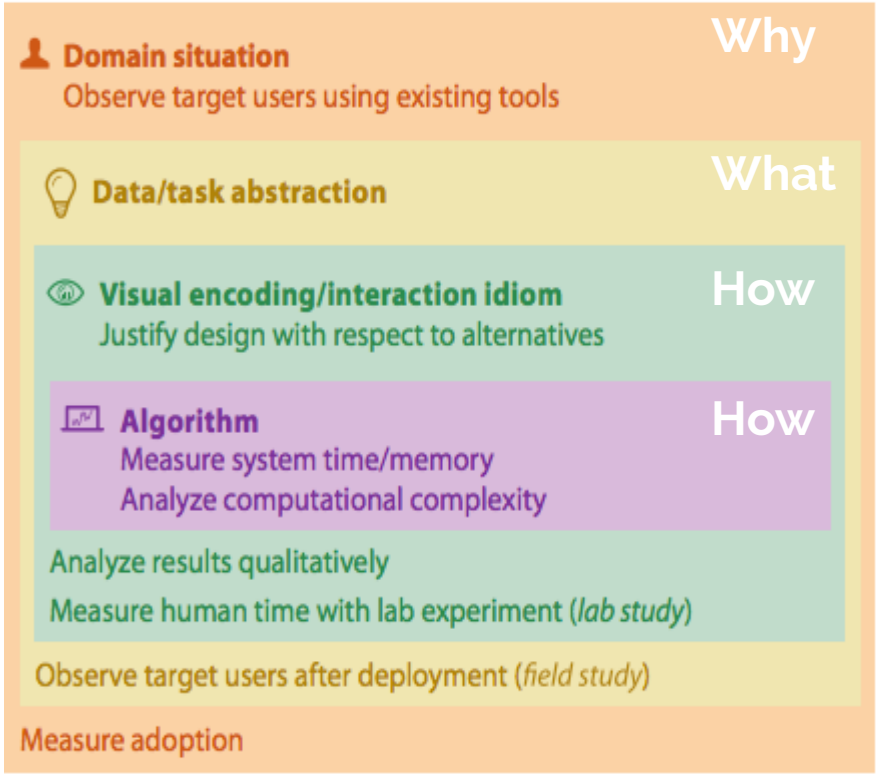
Methodology

Qualitative Methods, Domain Knowledge

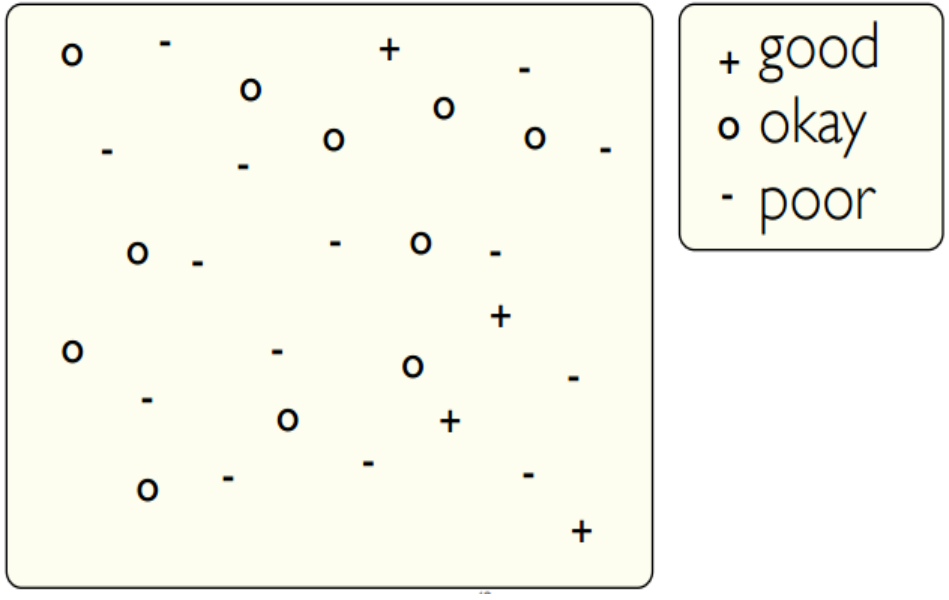
Design & Cognitive Science

Computer Science

Qualitative & Quantitative Methods



The "Design Space" metaphor

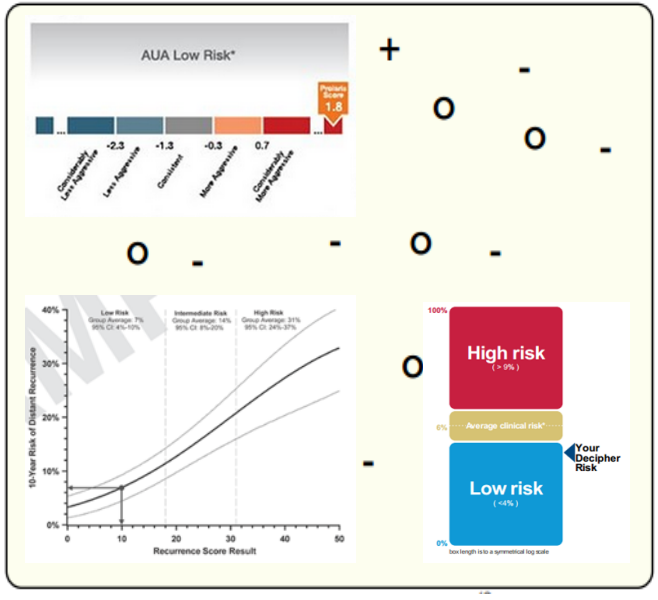


Sedlmair 2012
https://www.cs.ubc.ca/nest/imager/tr/2012/dsm/dsm-talk.pdf

How Data Visualization is like Statistical Modelling

OPTIMIZATION!

The "Design Space" metaphor

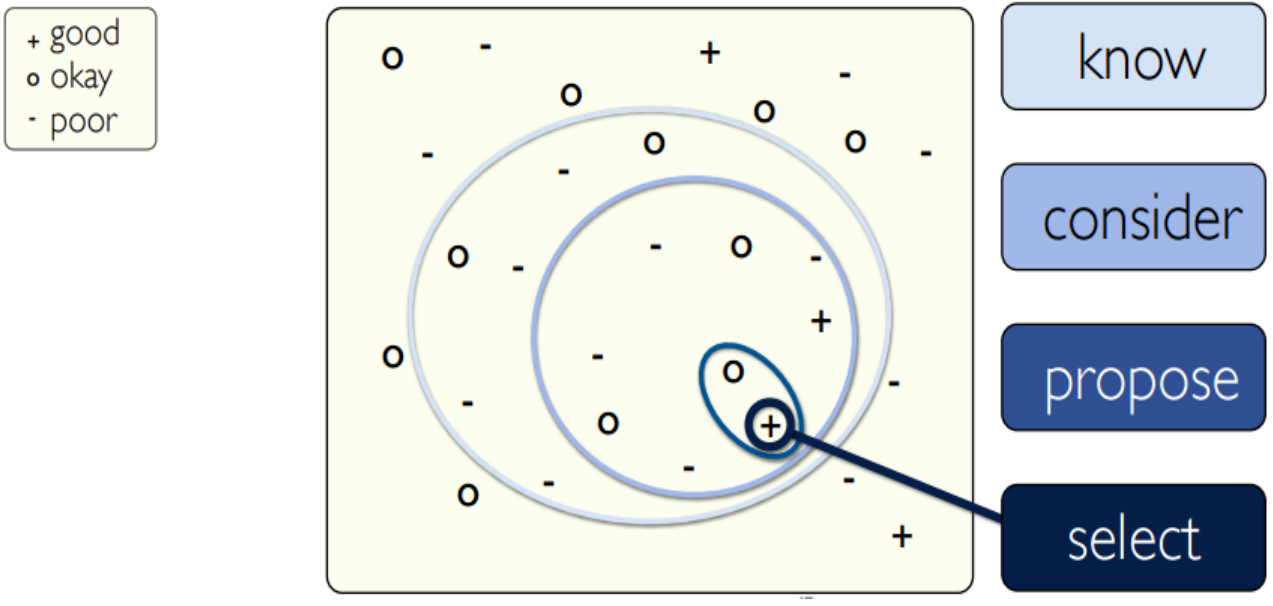


+ good
o okay
- poor

Sedlmair 2012
https://www.cs.ubc.ca/nest/imager/tr/2012/dsm/dsm-talk.pdf

Progressively Identify the Right Visualization

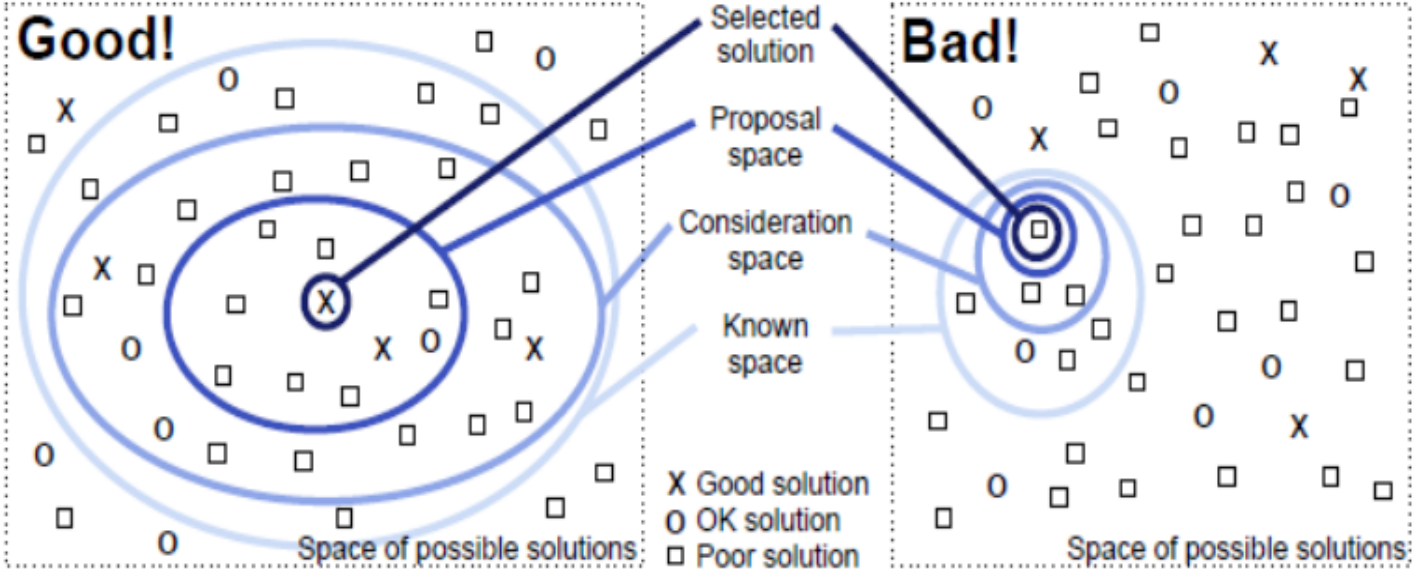
The "Design Space" metaphor



Use "why, what, and how" framework to guide the selection of the optimal design choice

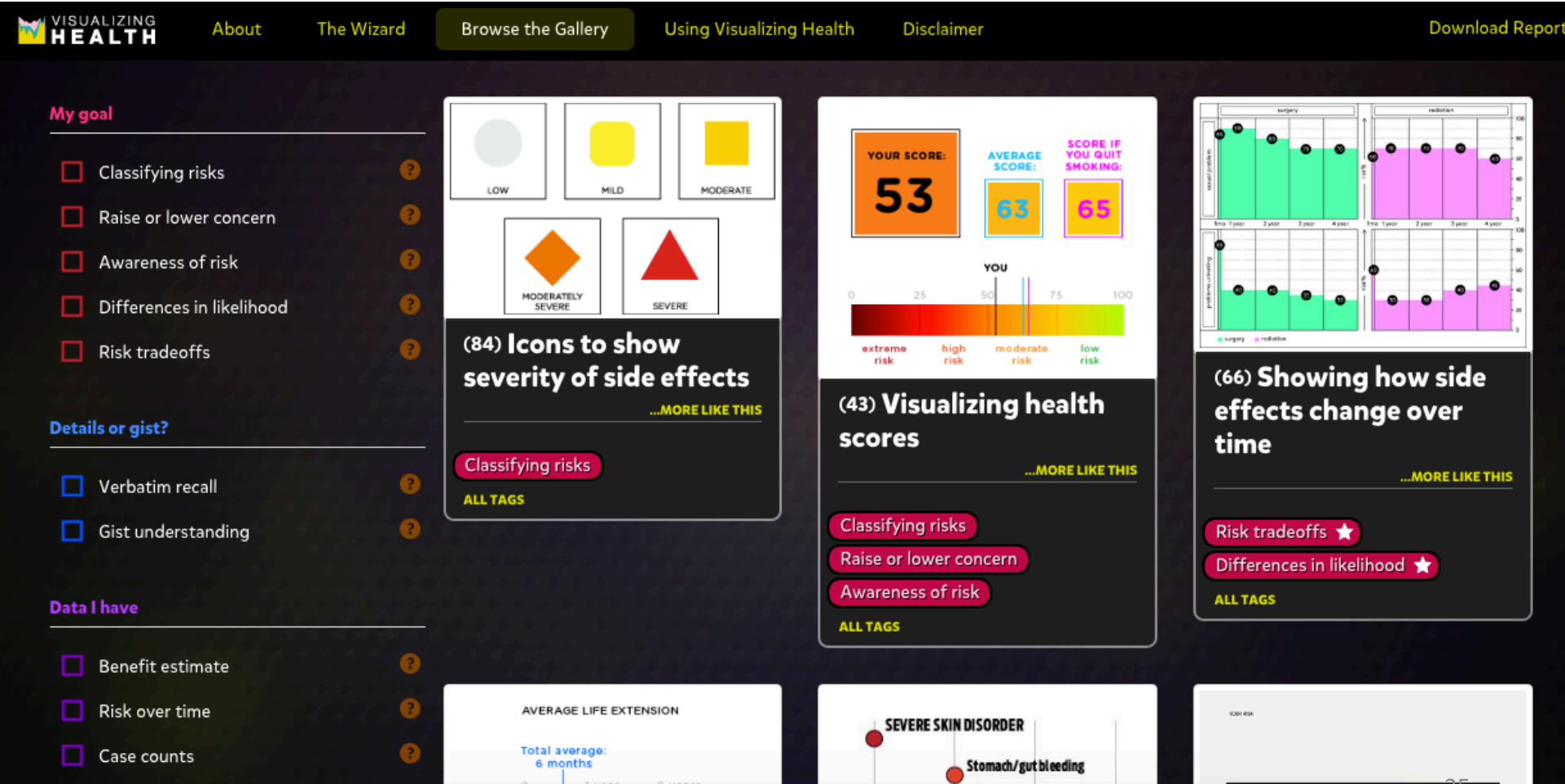
Sedlmair 2012
https://www.cs.ubc.ca/nest/imager/tr/2012/dsm/dsm-talk.pdf

The Importance of Thinking Broadly



Use "why, what, and how" framework to guide the selection of the optimal design choice

Designs for Visualizing Health Data (http://www.vizhealth.org/)



A preview of some things I am working on

BUT.....

How do we design good visualizations for **public health**?

Primary Research Question

To what extent and in what ways does the visualization of genomic, administrative, and contact network data support decision making for communicable disease prevention and control

Primary Research Question

To what extent and in what ways does the visualization of genomic, administrative, and contact network data support decision making for communicable disease prevention and control

aka. "How is visualization of communicable disease (public health) data useful? Can I quantify how useful it is?"

Some Example Sub Questions

What is the best way to ***visually represent data*** in an outbreak context to promote a rapid response?

How can stakeholders ***explore their data*** more effectively to identify areas of needs and develop effective outreach programs?

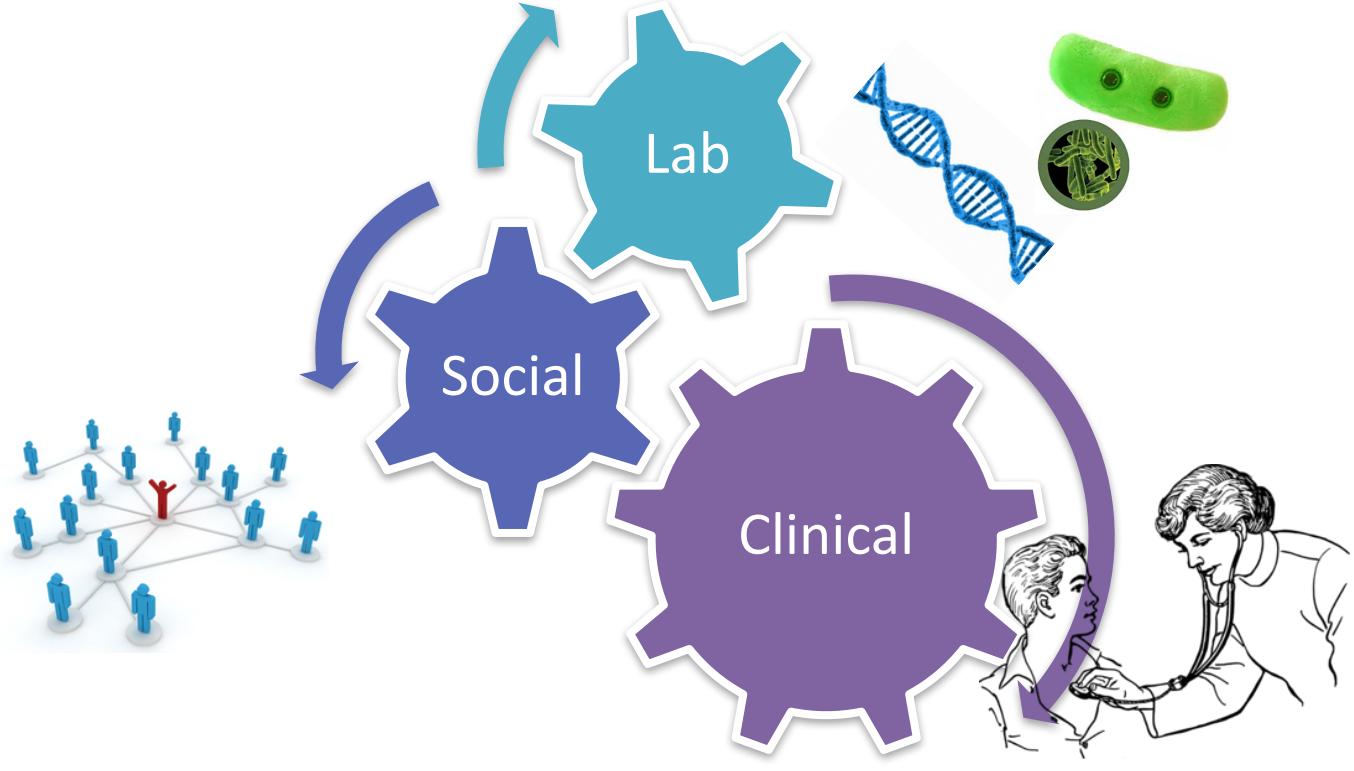
What is the most effective way to ***show*** genomic ***data over space and time***?

Example 1

Visualizing Tuberculosis data at the
British Columbia Centre for Disease Control

WHY

Combining Data will Prepare us for the Pandemics of the Future



But, that's a lot of data....

Can Visualizing TB data help Decision Support?

We wanted to create an **interactive** and **visual tool** that allowed our public health stakeholders to analyze the different data types

We want to understand how this tool can be used by different public health stakeholders

Medical Health Officers

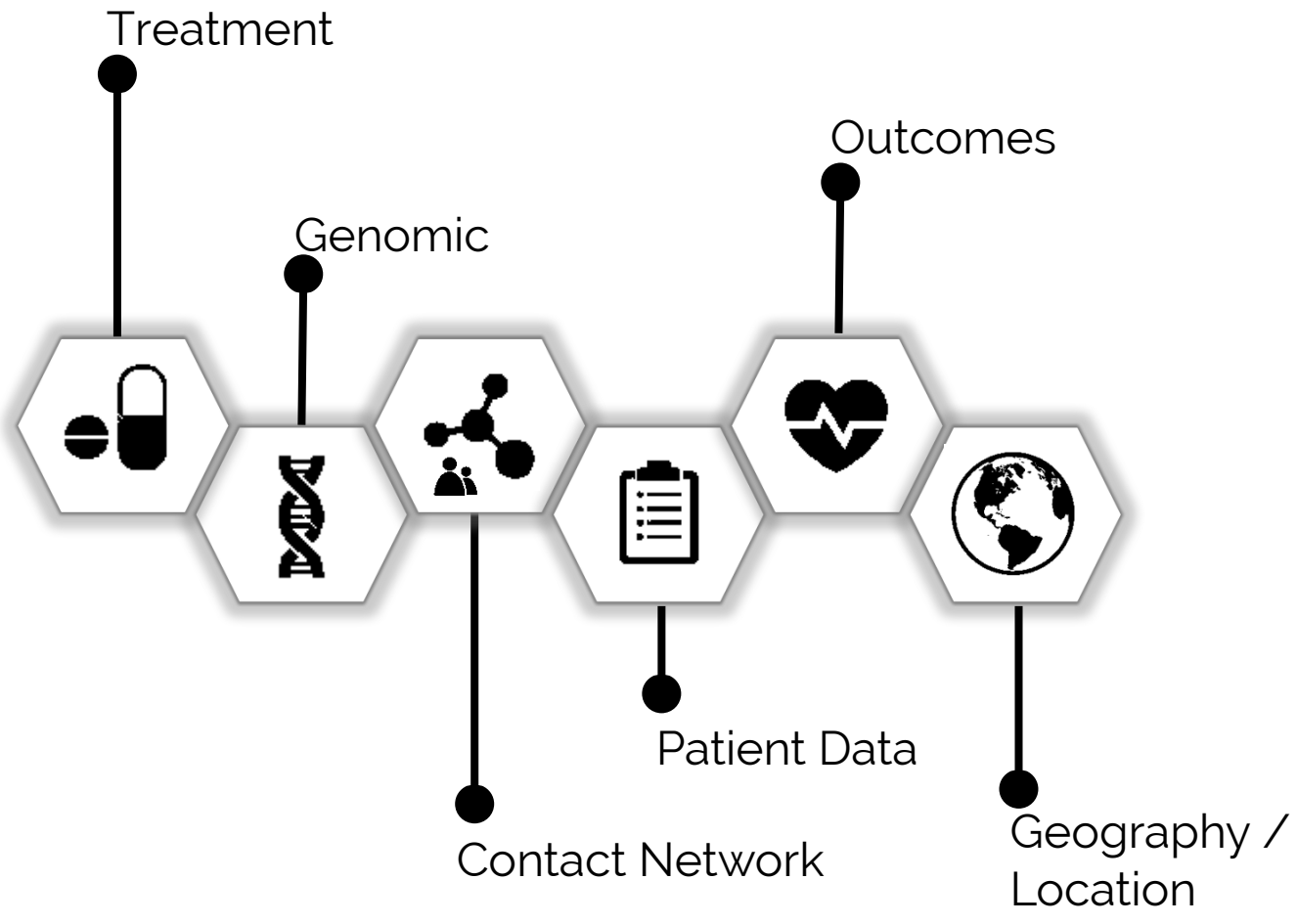
TB Clinicians

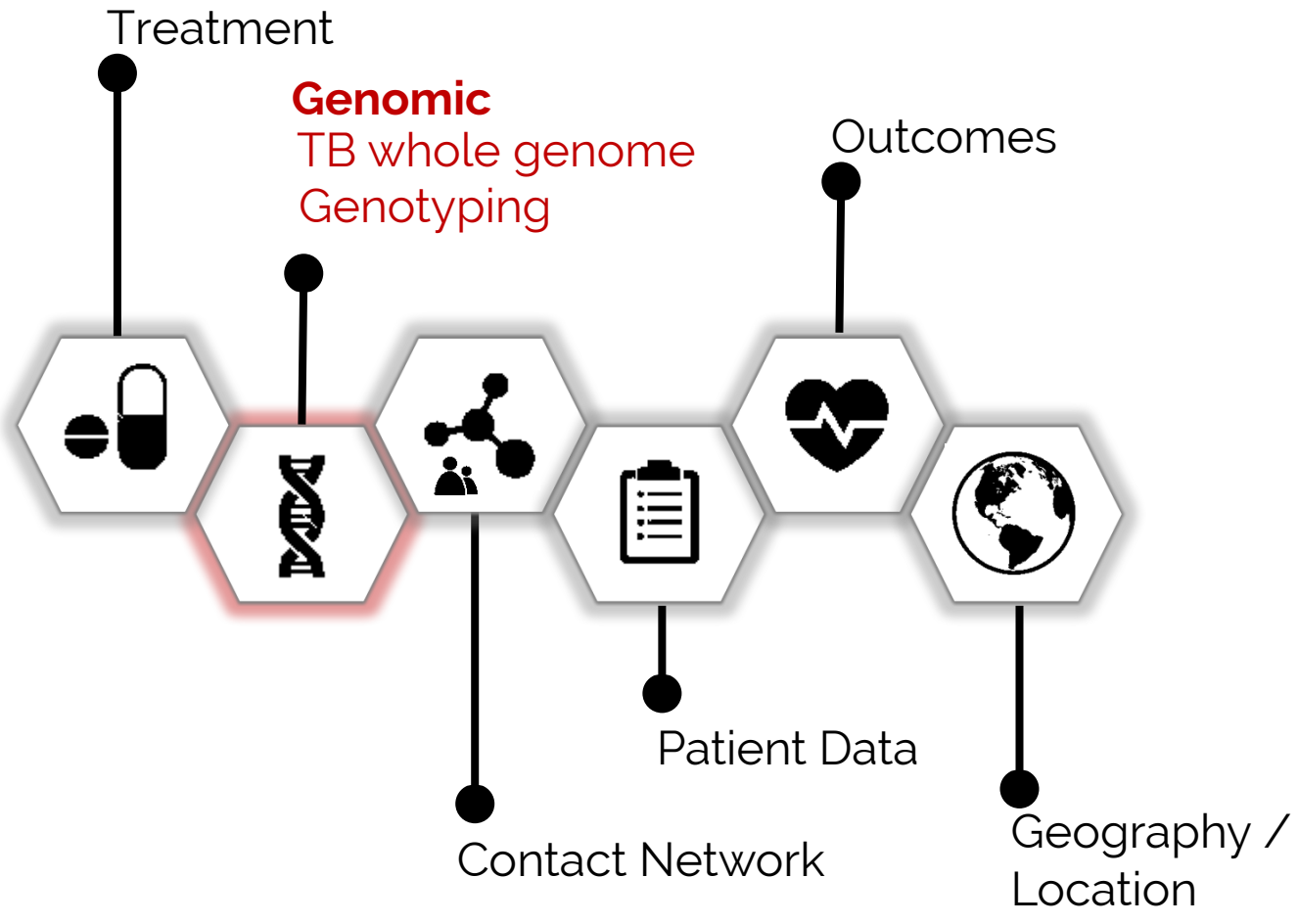
TB Nurses

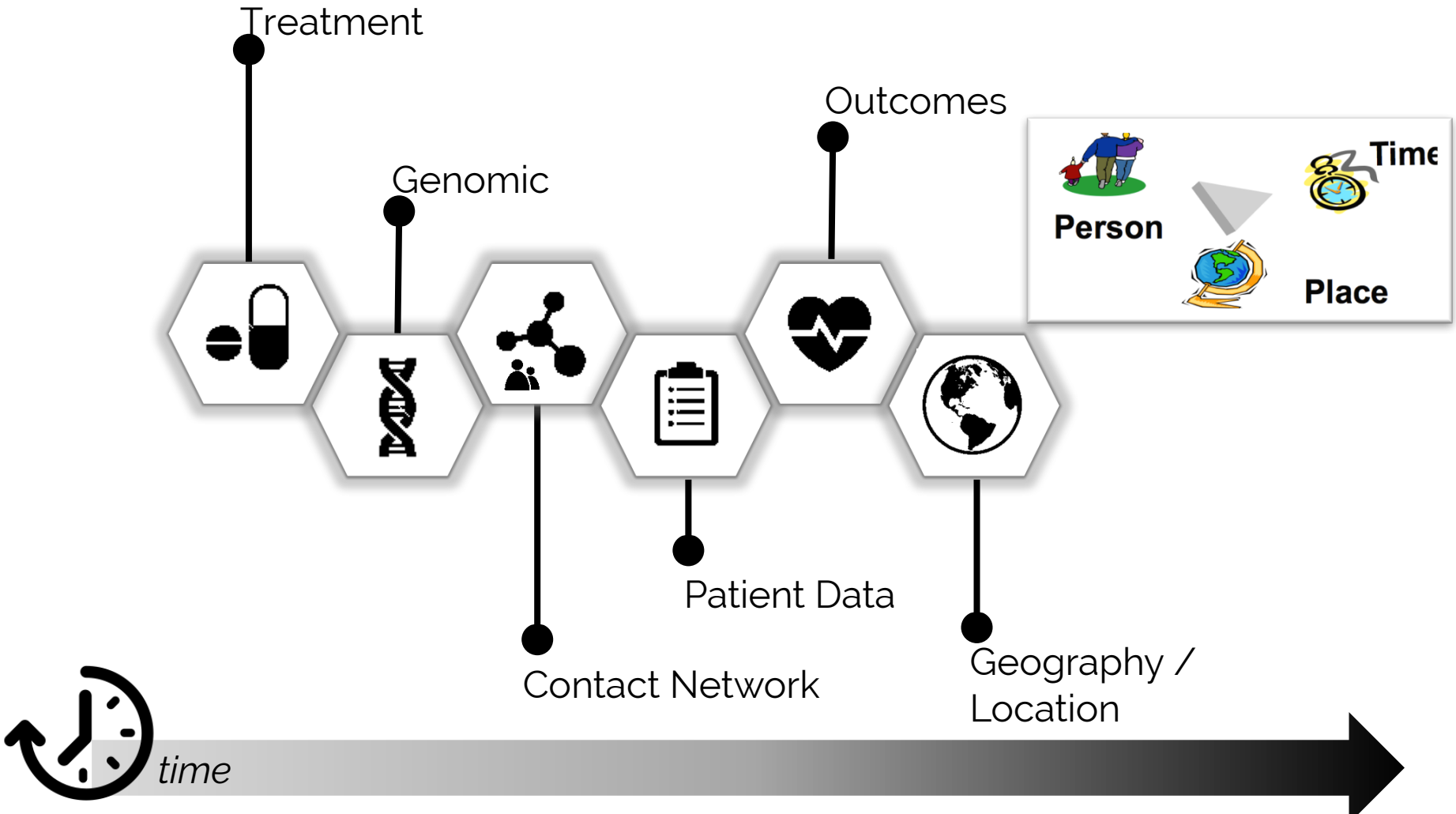
Researchers

Epis / Biostats

WHAT

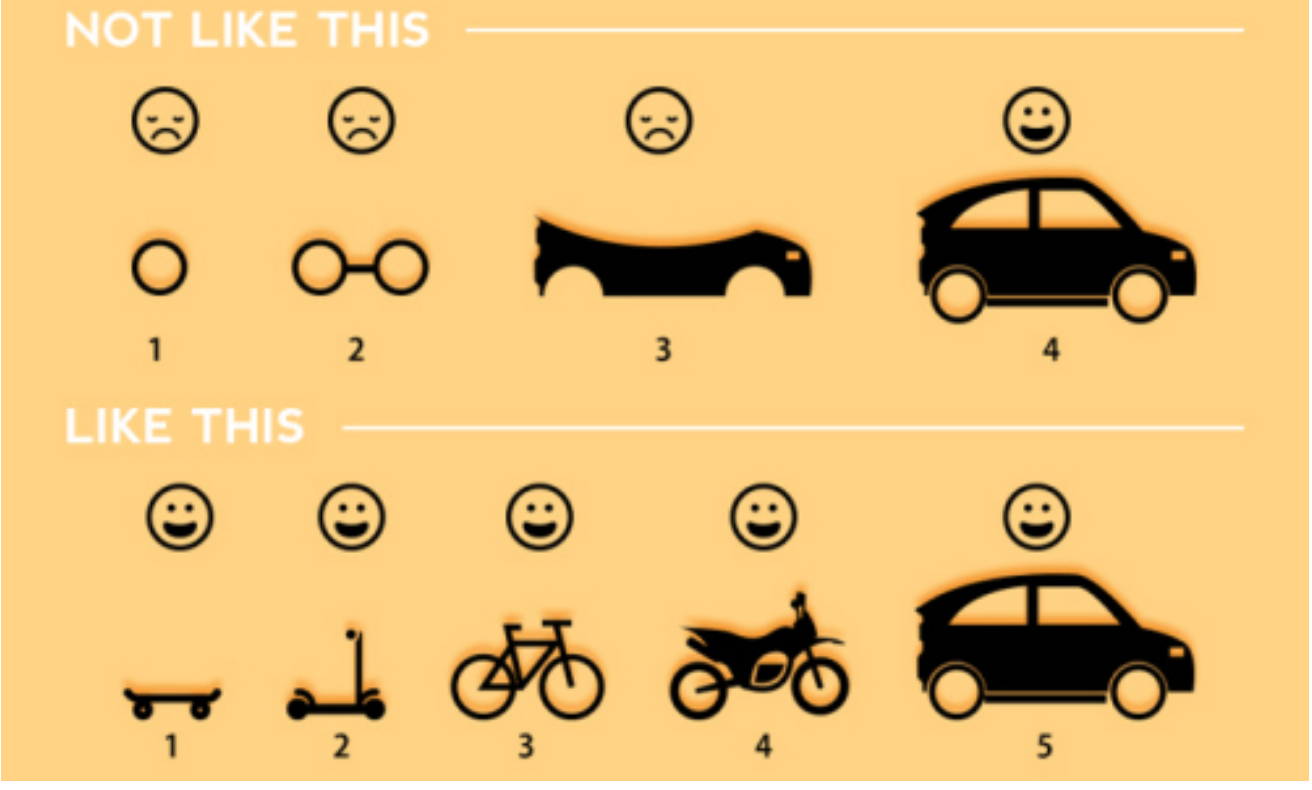

Treatment
Genomic
Outcomes
Patient Data
Contact Network
Geography / Location


Treatment
Genomic
TB whole genome
Genotyping
Outcomes
Contact Network
Patient Data
Geography /
Location


Treatment
Genomic
Outcomes
Person
Time
Place
Patient Data
Contact Network
Geography / Location
time

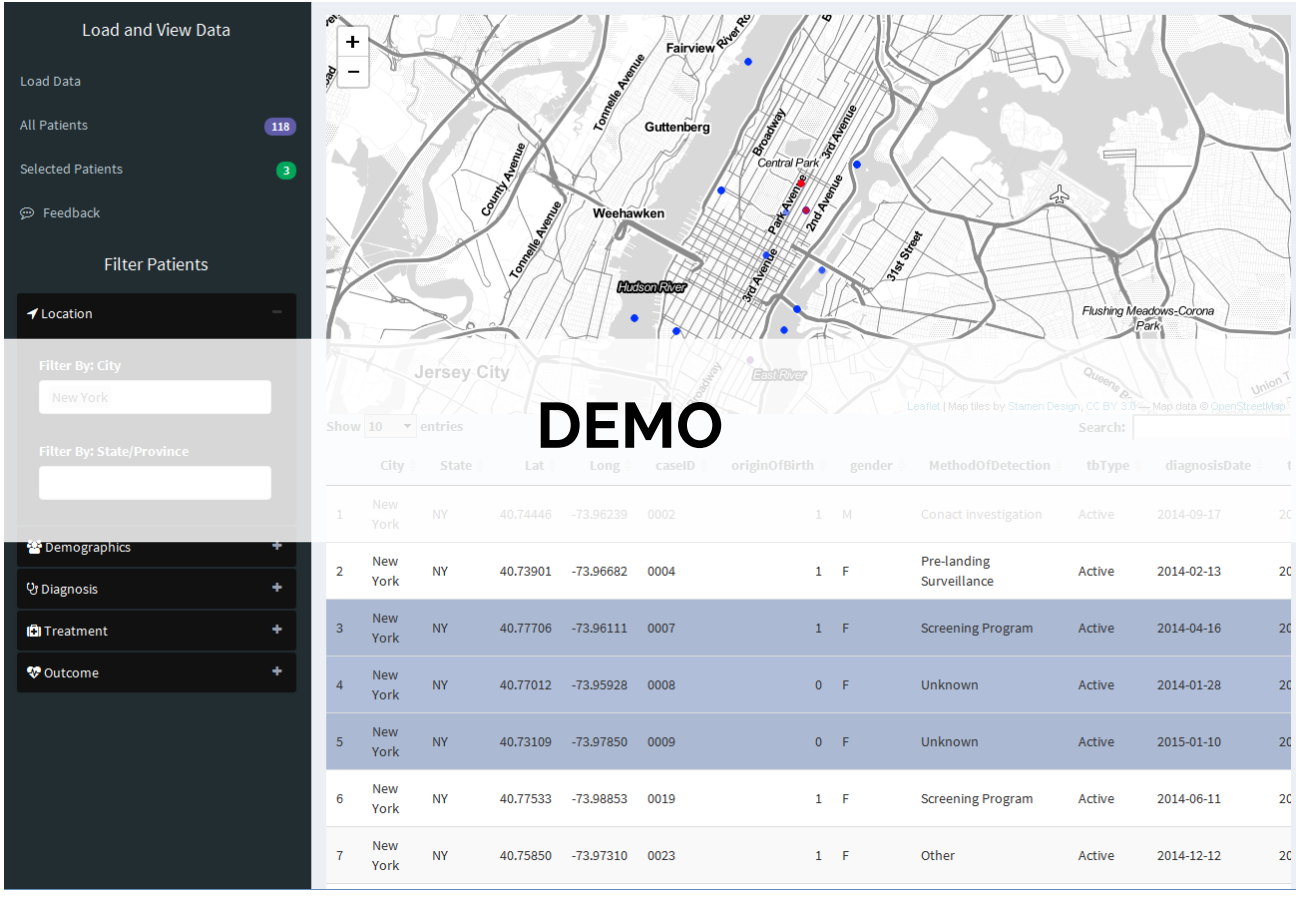
HOW

An Iterative Approach to Development



An iterative approach to development allows us to get feedback before committing to ineffective design choices

Introducing EpiCOGs



	City	State	Lat	Long	caseID	originOfBirth	gender	MethodOfDetection	tbType	diagnosisDate	t
1	New York	NY	40.74446	-73.96239	0002	1	M	Conact investigation	Active	2014-09-17	20
2	New York	NY	40.73901	-73.96682	0004	1	F	Pre-landing Surveillance	Active	2014-02-13	20
3	New York	NY	40.77706	-73.96111	0007	1	F	Screening Program	Active	2014-04-16	20
4	New York	NY	40.77012	-73.95928	0008	0	F	Unknown	Active	2014-01-28	20
5	New York	NY	40.73109	-73.97850	0009	0	F	Unknown	Active	2015-01-10	20
6	New York	NY	40.77533	-73.98853	0019	1	F	Screening Program	Active	2014-06-11	20
7	New York	NY	40.75850	-73.97310	0023	1	F	Other	Active	2014-12-12	20

EpiCogs is a data viewer and currently a sandbox environment for developing data visualizations

Factors Influencing the Current Design

Needs of individuals

Gathered through meetings, dialogue with individuals, and various iterations of EpiCOGs

Technology Changes

Support for data visualization tools in R improved greatly allowing for the creation of better data visualizations

Data Driven Interface and Analysis

Created a data driven interface that is responsive to the user's data.

Policies and Procedures

Existing policies and procedures at the BCCDC inform the utility of such a tool and how it can integrate into existing workflows

Initial Work & Next Directions

Much initial work was to understand the tool's feasibility

Could it meet the needs of stakeholders?
How could it integrate (security & workflow)?
How could it be supported long term? (Choice of R)
Could we build a useful tool in R?

Next phases will explore genotypes, genomics, and contact networks

Right now, users can filter based on assigned genotype clusters (which will show patients on map), but we're working towards better visual and interactive design for these data

This is an Open Source Project

TRY THE DEMO:

https://amcrisan.shinyapps.io/EpiCOGSDEMO/

GET THE CODE

(& contribute to the project!) :

https://github.com/amcrisan/EpiCOGS/

Call for Guinea Pigs!

To make relevant tools I need feedback!
If you want to be involved and get project updates let me know!

E-mail: anamaria.crisan@bccdc.ca

Twitter: @amcrisan

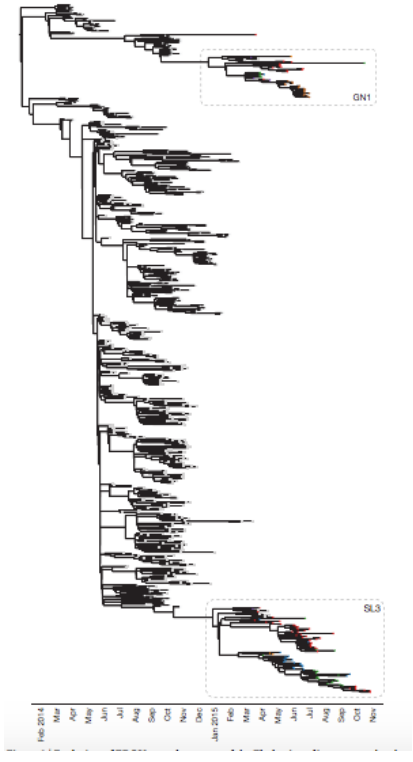
Web : cs.ubc.ca/~acrisan

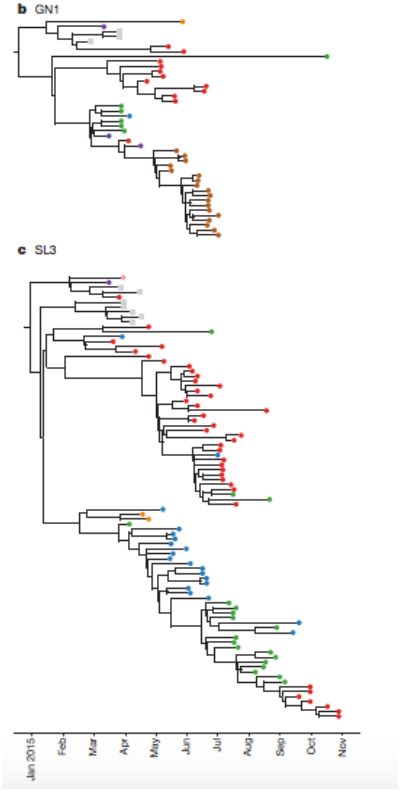
Example 2

Visualizing the Ebola Outbreak – An example of a design process

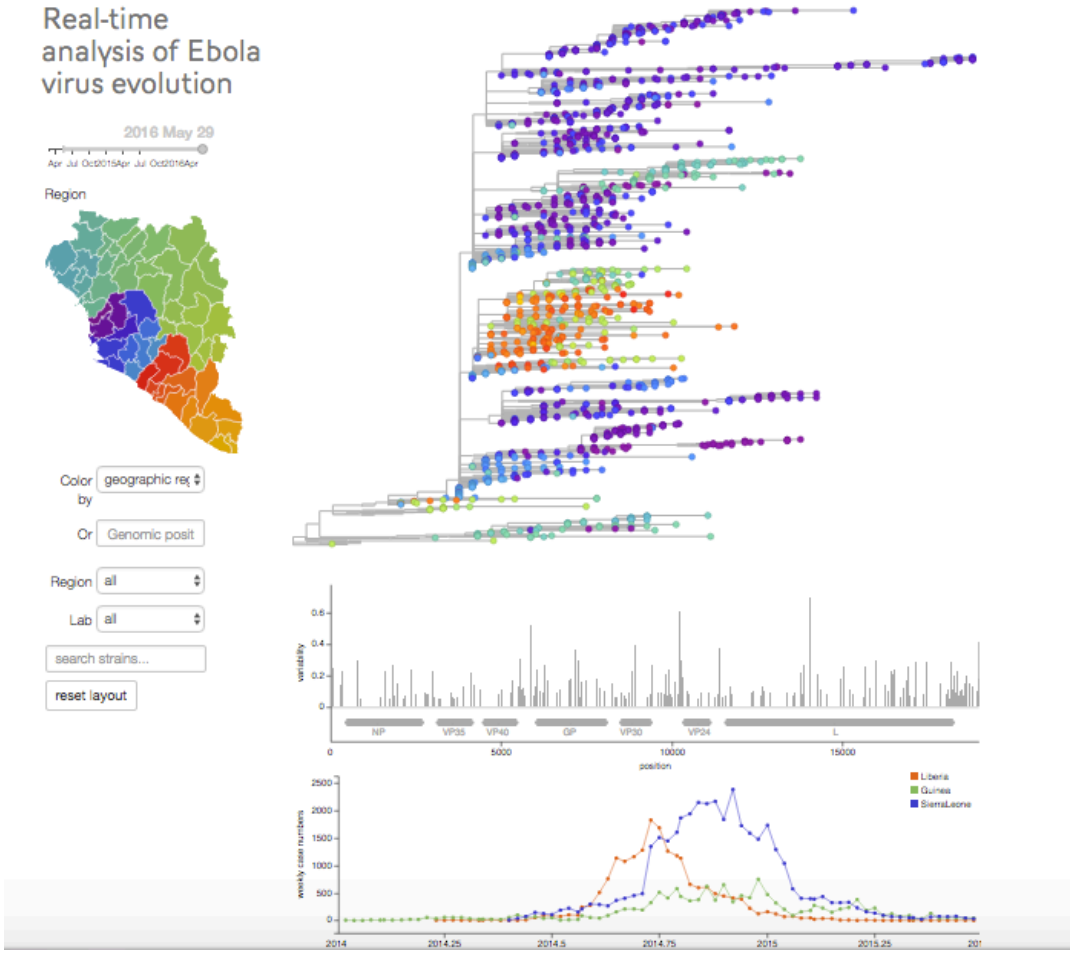
This was what we started with

A very familiar layout, all the information is there, but you have to do some work to put it together





This was what we started with



Bedford Lab – Next Strain

Can we improve the Design of the Visualization?

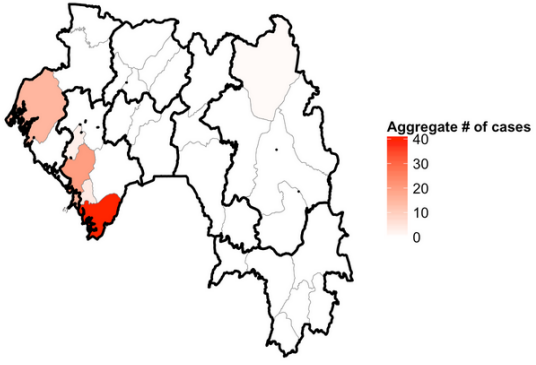
- How do different public health stakeholders use the data?

- Epidemiologists want to know:
  - Where is it spreading?
  - How is it spreading?
  - How many people are impacted?

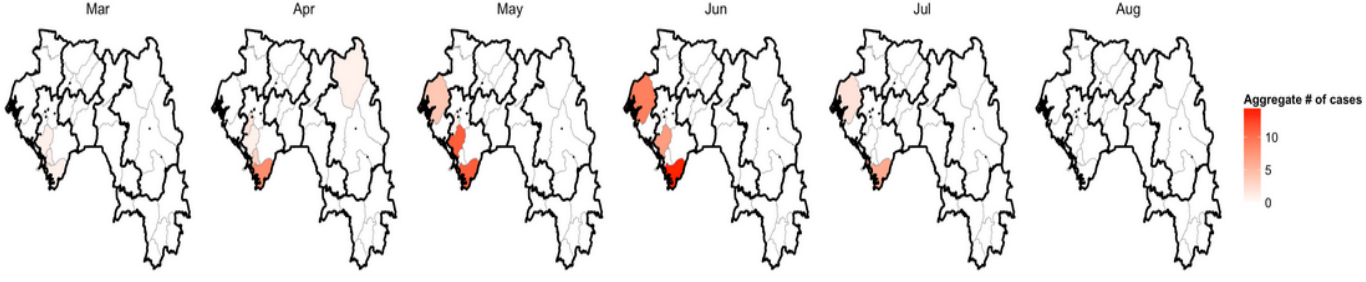
- Researchers want to know:
  - What's spreading?
  - How similar are the outbreak clusters?
  - How is changing over time?

Step 1: Small multiples by time

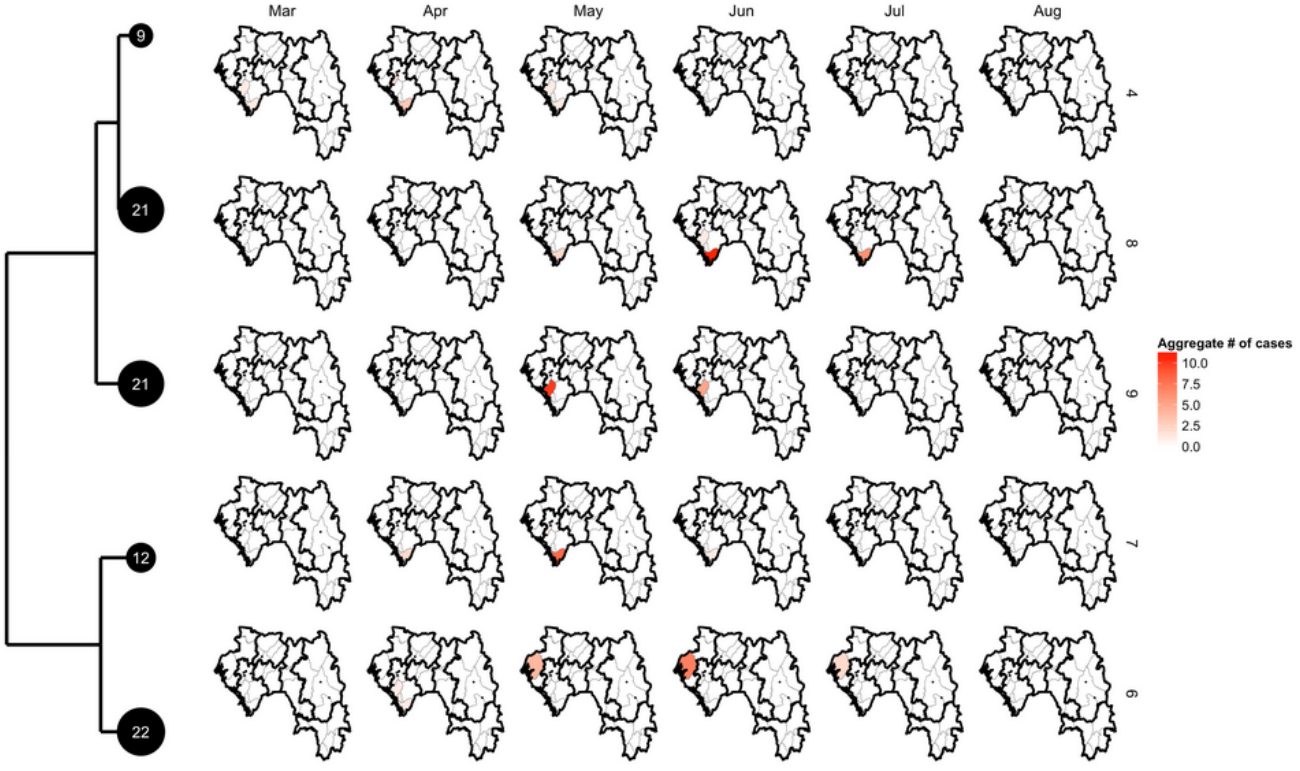
Aggregate case distribution over entire sampling period



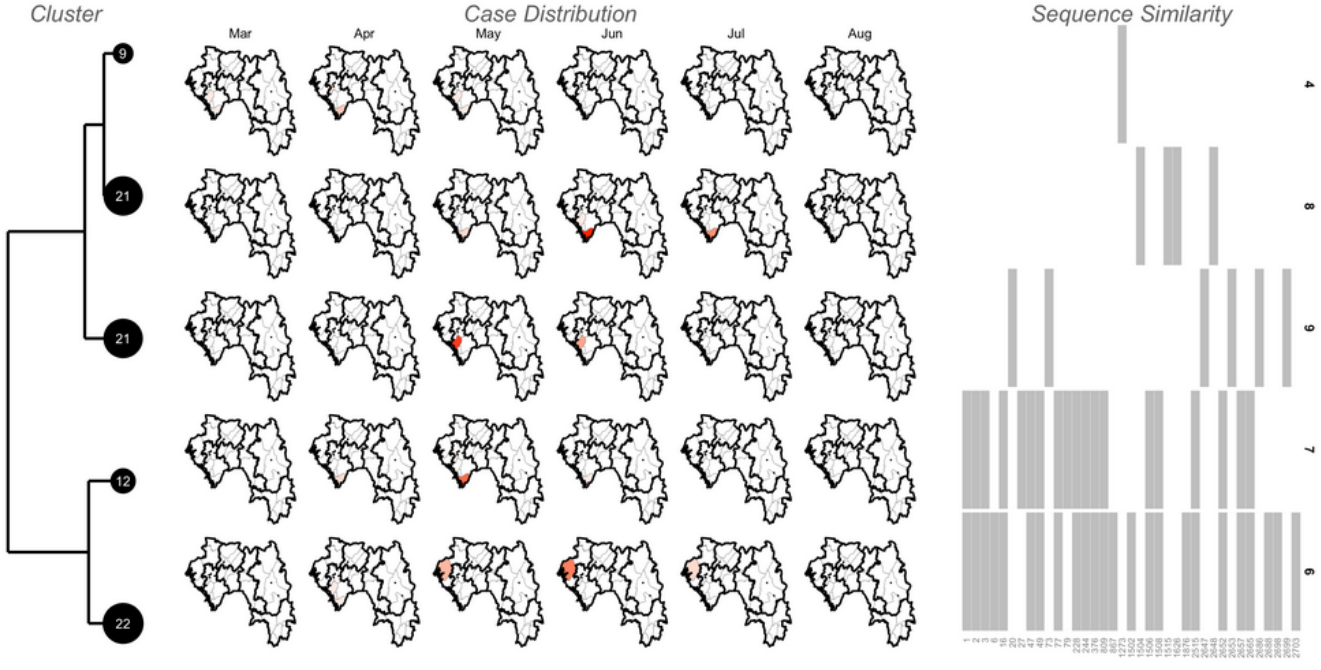
Aggregate case distribution by month



Step 2: Small multiples by time and genome cluster

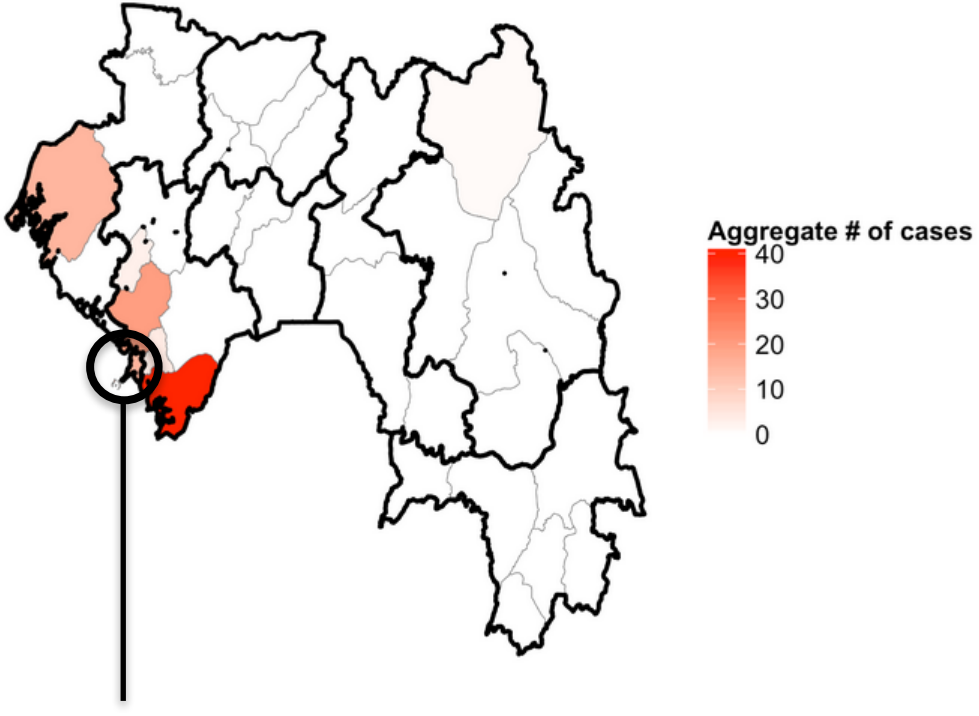


Step 3: Small multiples by time and genome cluster and with sequence similarity



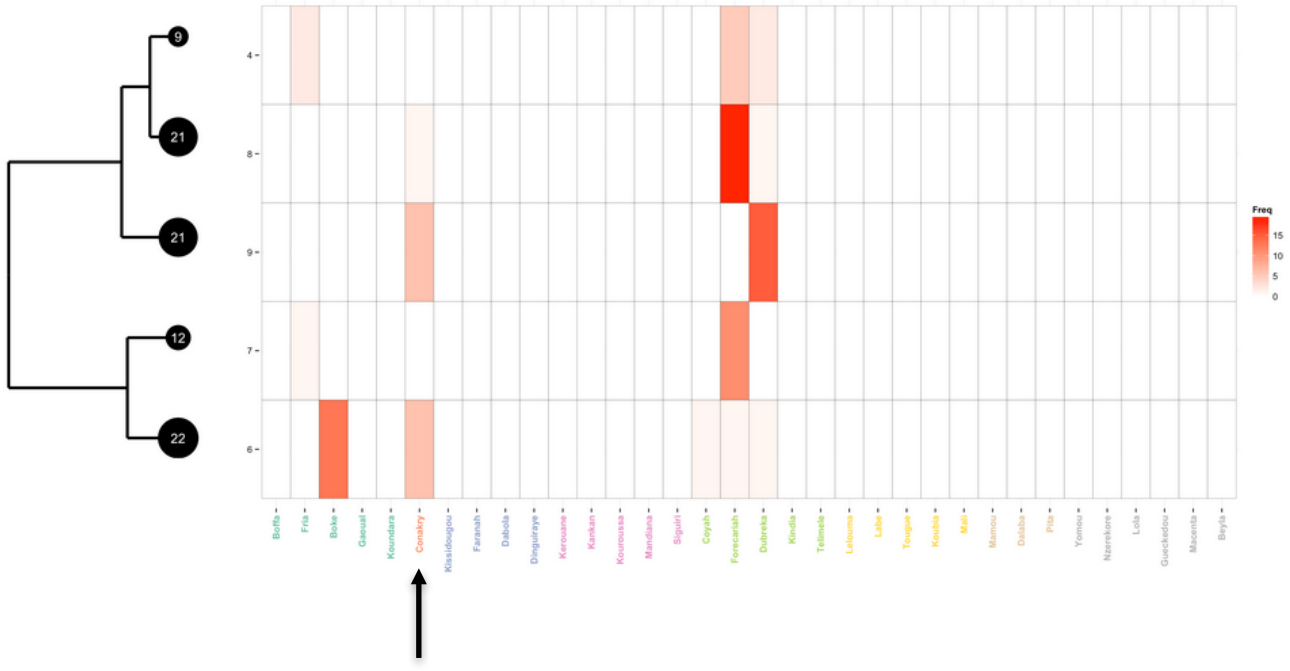
White: dominant nucleotide
Grey : less dominant nucleotide

By abstracting the geography, we can represent more data more easily

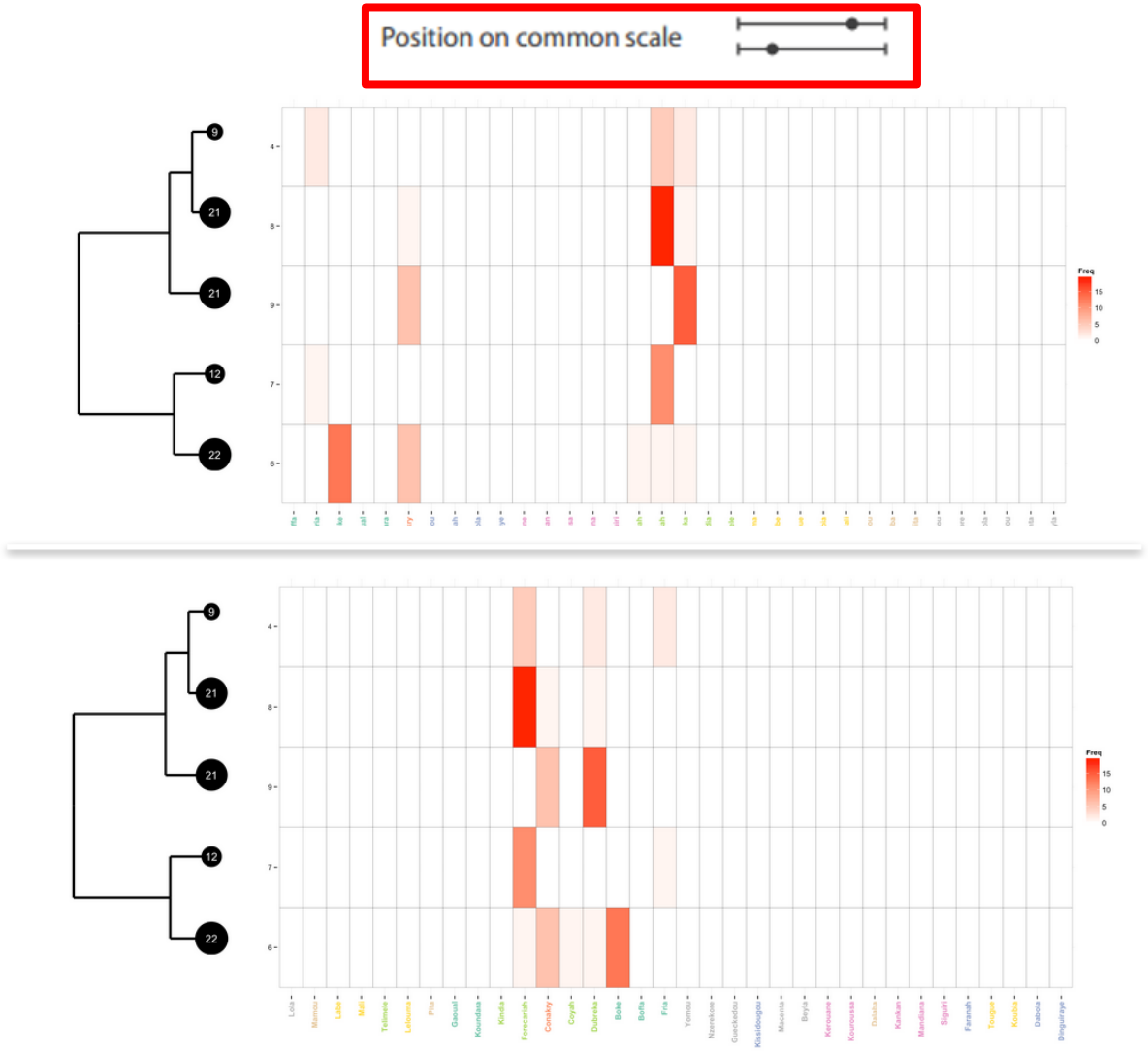


Highly populous capital is very difficult to see

By abstracting the geography, we can represent more data more easily



Capital city gets a more prominent view



X-axis ordering

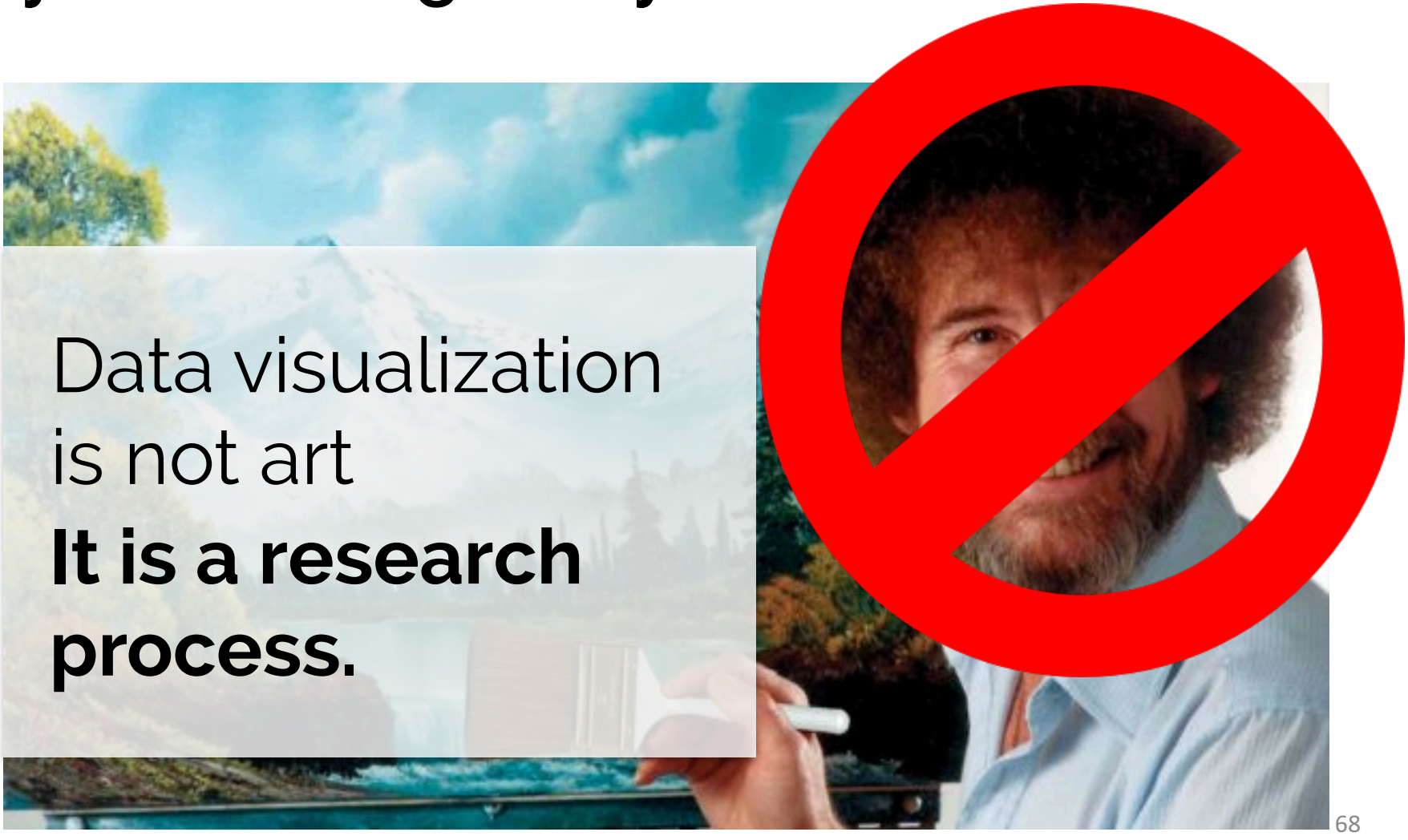
Alphabetically within high level administration regions

Geographic distance

Part 3:

Take home messages

Beyond Building Pretty & Cool Visualizations



Take Home Messages

Data Visualization is not an art or graphic design project

Relevance (utility) and usability trump aesthetics

Take Home Messages

Data Visualization is not an art or graphic design project

Relevance (utility) and usability trump aesthetics

Deciding upon the most appropriate data visualization can be a research problem

Design & Evaluation

Think about "why, what, and how" framework

Parallels to finding the right statistical model

Take Home Messages

Data Visualization is not an art or graphic design project

Relevance (utility) and usability trump aesthetics

Deciding upon the most appropriate data visualization can be a research problem

Design & Evaluation

Think about "why, what, and how" framework

Parallels to finding the right statistical model

Think broadly, progressively find the right data visualization

The Design Space Concept

Iterative development

This would work not be possible without these fine people

The Gardy Lab
Dr. Jennifer Gardy
Jennifer Guthrie

PHSA Reference Laboratory
Dr. Patrick Tang
Hope Lapointe
Clare Kong

BCCDC CDPACS
Ciaran Aiken
Laura MacDougall
Mike Coss
Sunny Mak
Mike Otterstatter
Robert Balshaw

UBC Computer Science
Dr. Tamara Munzner
The *InfoVis* group

BCCDC Clinical TB Clinical Team
Clinicians
Dr. Maureen Mayhew
Dr. James Johnston
Dr. Jason Wong (CPS)
Dr. Victoria Cook

Nurses
Nash Dhalla
Michelle Mesaros

Epidemiologists
Dr. David Roth

The large team of individual's from BC's HAs and HSDAs without whom there would be no data.